IBIMA Publishing
Communications of the IBIMA
https://ibimapublishing.com/articles/CIBIMA/2024/176992/
Vol. 2024 (2024), Article ID 176992, 14 pages, ISSN: 1943-7765
https://doi.org/10.5171/2024.176992

Research Article

Enhanced BERT Approach to Score Arabic Essay's Relevance to the Prompt

Rim AROUA MACHHOUT And Chiraz BEN OTHMANE ZRIBI

National School of Computer Sciences, Manouba University, Tunisia

Correspondence should be addressed to: Rim AROUA MACHHOUT; rim.aroua@ensi-uma.tn

Received date:31 October 2023; Accepted date:2 February 2024; Published date: 1st March 2024

Academic Editor: Chiraz El Hog

Abstract

In recent years, automated essay scoring systems have seen significant progress, particularly with the integration of deep learning algorithms. This shift marks a move away from the traditional focus on style and grammar to a more in-depth analysis of text content. Despite these advancements, there remains a limited exploration of the essay's relevance to the prompts, especially in the context of the Arabic language. In response to this lack, we propose a novel approach for scoring the relevance between essays and prompts. Specifically, our aim is to assign a score reflecting the degree of adequacy of the student's long answer to the open-ended question. Our Arabic-language proposal builds upon AraBERT, the Arabic version of BERT, and enhanced with specially developed handcrafted features. On a positive note, our approach yielded promising results, showing a correlation rate of 0.88 with human scores.

Keywords: Automated Essay Scoring systems, Enhanced BERT with handcrafted features, Relevance to the prompt, Arabic language.

Introduction

An automated essay scoring system (AES) aims to automate the evaluation of students' long response to open-ended questions, thereby reducing the burden on teachers in terms of time and effort. It also helps to make distance learning more effective as stated by Mizumoto et al (2023). To ensure maximum reliability, these systems try to simulate the evaluation process followed by humans. Thus, essays must be assessed on all levels: style, lexical, semantic, and contextual. However, contextual analysis of a long text remains the most challenging task, especially when dealing with a language as difficult to process, such as Arabic.

Recently, proposals based on machine learning and deep learning have shown promising advances in terms of results. Nevertheless, despite progress in this field, a literature review by Ramesh et al (2022) reveals that current AES systems face significant challenges when it comes to effectively analyzing the contextual subtleties within essays. These obstacles include the need to

Cite this Article as: Rim AROUA MACHHOUT And Chiraz BEN OTHMANE ZRIBI (2024)," Enhanced BERT Approach to Score Arabic Essay's Relevance to the Prompt ", Communications of the IBIMA, Vol. 2024 (2024), Article ID 176992, https://doi.org/10.5171/2024.176992

analyze coherence and cohesion in evaluation, the development of ideas evoked by the student and the relevance of the essay's content to the prompt. Similarly, when addressing the Arabic language, proposals in this area remain relatively limited.

In this context, we intend to present a new approach aimed at evaluating the relevance of essays written in Arabic language and produced by primary school students in response to a given prompt. In fact, we mean by "evaluation of the relevance of the essay to the prompt" the scoring of how well the essay's content addresses the question's requirements.

Our proposal is based on the use of the BERT deep learning algorithm, which is recognized for its performance in this field but so far not well exploited for the Arabic language. Additionally, we aim to enhance BERT with handcrafted features. These especially developed features have emerged from multiple contextual analyses associated with the concept of 'relevance'.

The remainder of the paper is organized as follows: in section 2, we present some related works. We then delve into our proposed methodology in section 3, followed by our experiments and results in section 4. We conclude with a discussion and future works in the last section.

Related Works

Recent research in the field of Automated Essay Scoring systems has made significant progress. This proposal deals more deeply with the contextual characteristics of the essays, as discussed in the review by Jong et al (2023). Among the criteria examined, coherence was the most frequently studied. This concept is explored in different works, including Zupanc et al (2014), Li et al (2018), Farag et al (2018), Tay et al. (2018), Palma et al (2018) and Ramesh et al (2022). In contrast, Crossley et al (2013) and Salim et al (2019) focus more on cohesion. Additionally, Persing et al (2014) present an essay scoring system that considers the adherence of the content to the prompt. However, according to Ramesh et al (2022), while AES systems have progressed in considering these contextual criteria, there are still challenges to achieve a rating as competitive as that performed by an experienced human evaluator.

For the Arabic language, research in this direction is still late, especially as few studies have used deep learning and, in general, the scoring of essays is based on features as outlined by Machhout et al (2021). Thus, Alqahtani et al (2019) proposed a rule-based system that defines an outline based on evaluation criteria inspired by Arabic literary resources and the experiences of university teachers. The criteria include spelling, grammar, structure, cohesion, style, and punctuation. The advantage of this study is that it takes into account two contextual criteria, namely coherence and style. The average accuracy of this proposal was 73%. However, we noticed that they assess the 'relevance' of the essay to the prompt, referred to as "coherence". They use cosine similarity for this purpose; however, it is important to note that this measure alone cannot be considered significant in this context. For example, if a student responds by rewriting the prompt one or more times, in this case the similarity will be high, and the answer is considered 'relevant" then noted as well consistent.

In their work, Azmi et al (2019) present an approach based on latent semantic analysis (LSA), Rhetorical structure theory (RST) and other features. What is notably distinctive about this approach is their focus on the essay's contextual level, which is concretely reflected in the criterion they call 'writing style'. This includes various criteria such as essay cohesion, checking for duplicate sentences, the presence of vernacular terms, and the total length of the essay. However, they did not assess the relevance of the response to the prompt. According to the results presented in the article, this approach achieves a correlation of 0.759 with teachers' scoring.

Alqahtani et al (2020) propose an approach based on support vector regression (SVR). They create specific models for each evaluation criterion, i.e., spelling, structural coherence, style, and punctuation. These models use different types of features. The final essay evaluation is obtained by combining the results of different individual models. This approach involves the essay's context by assessing its structure and coherence. The evaluation of coherence entails measuring how well the essay's parts are related to the title and the cohesion between the parts of the essay. As a result, it achieves a notable increase in the correlation rate with expert evaluation, reaching a score of 0.87.

Another recent proposal in Arabic is presented by Alobed et al (2021). This approach is based on the support vector machine (SVM) and Arabic

wordnet. This proposal focuses on the semantic analysis of essays and is via the integration of Arabic Wordnet. However, the authors did not detail the results.

The Essay's Relevance to the Prompt

The analysis of previous works targeting AES systems has shown that several studies did not consider the concept of 'relevance to the prompt'. In the case where it was addressed, its evaluation has been limited to using cosine similarity between the essay and the prompt as in Alobed et al.'s work (2021), or it was regarded as an element of 'coherence' as explored in the study by Alqahtani et al (2019).

Taking inspiration from previous works and considering their limitations, this article presents a novel approach for scoring 'the essay's relevance to the prompt" within the context of automated essay scoring systems in Arabic language. We propose an approach based on the BERT algorithm and enhanced with handcrafted features.

In fact, the deep learning algorithm BERT has demonstrated its performance in terms of contextual text processing. This algorithm is trained on a large corpus and can be used directly or fine-tuned on a new corpus, even of small size. BERT's key strength lies in its ability to represent words within their contexts. Hence, the same word can have different representations depending on its context of use. For the Arabic language, diverse versions trained on Arabic corpora are available. Notable examples include the proposals by Antoun et al (2020), Safaya et al (2020), Abdul-Mageed et al (2020), and Inoue (2021). All these advantages encouraged us to choose BERT.

On the other hand, to ensure that our work doesn't only remain theoretical, but rather is performed usefully in practice, the selection of handcrafted features was not arbitrary. We based on the criteria set by the Tunisian Ministry of Education (Fig 1). Noteworthy, the assessment of relevance ("الملاءمة" in Arabic) is a fundamental criterion in the evaluation of essays. The starting point for evaluating an essay is whether a response (essay) conforms to what it was asked to do (prompt). If a student's response diverges significantly from the assigned topic, further evaluation of other levels becomes unnecessary. Therefore, we processed to investigate some primary education teachers to discern the essential points to consider when assessing an essay's relevance. Three key criteria emerged from this discussion: the compatibility between the requested theme and that addressed in the response, the respect of the type of text requested (narration, description, or dialogue) and the sharing of some keywords. We will detail these features in the following sections.

It should be highlighted that essays are presented as long texts in response to open-ended questions. Moreover, our approach supports open prompts and is not restricted to a single prompt or domain. Refer to Figure 1 for several examples of prompts considered in our work.

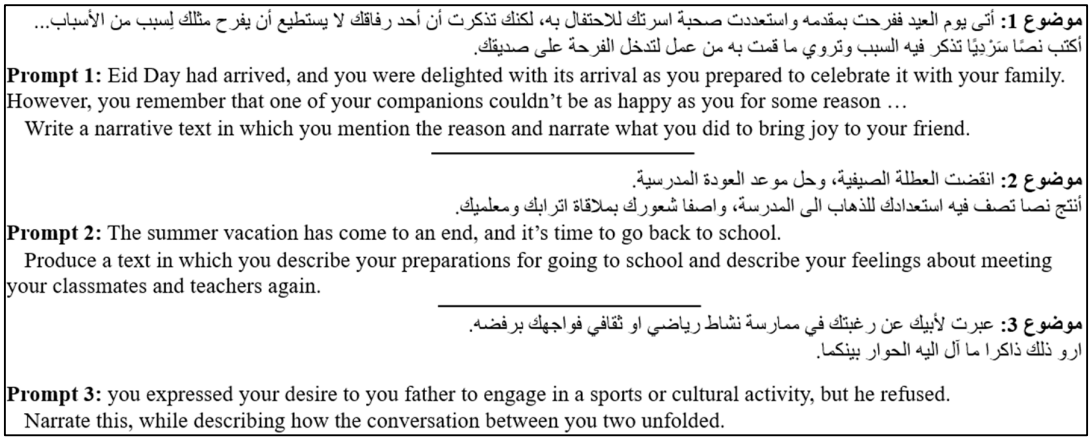

Fig 1. Prompts examples from our dataset

Proposed "BertRelevance" Architecture

Our idea is to develop a system called "BertRelevance", extending the architecture of BERT introduced by Devlin et al (2018). We enhance BERT with handcrafted features and incorporate on top additional stacked Multi-Layers Perceptron (MLP), taking inspiration from Gu et al (2021). Bert is tasked with analyzing the contextual representation of inputs, while the MLP are responsible for predicting the appropriate score. Since, the Arabic-language version of BERT is trained on relatively small corpora, we decided to strengthen our model's learning by incorporating handcrafted features. However, it should be noted that, regarding Devlin et al (2018), a shortcoming of BERT lies in its limited input capacity of 512 words. Given that our task involves the manipulation of long texts, the simple concatenation of the essay and prompt in the system input as done by Beseiso et all (2021) is impossible. To overcome this limitation, we adopted a two-step approach, as follows:

- **Features extraction:** In this step, we extract the contextual representation of the prompt based on AraBERT, an Arabic-language version of BERT proposed by Antoun et al (2020). This vector is denoted as **'Prompt_Emb'**. Then, from the essay and the prompt, we extract various features to form a vector of features called **'Vector_Features',** which we will describe in more detail later in Section 3.2.

- **Score prediction:** The essay is initially analyzed by BERT, in our case AraBERT, the Arabic-language version. The token [cls] obtained encapsulates the contextual representation of the essay. At this stage, we integrate the two previously prepared vectors, **'Prompt_Emb'** and **'Vector_Features'** resulting in the **"Combined_Vector"** as shown in Figure 4. This will be transmitted to the second part of our system, namely the MLP, which are responsible for predicting the appropriate score. Figure 2 illustrates the overall architecture of the system

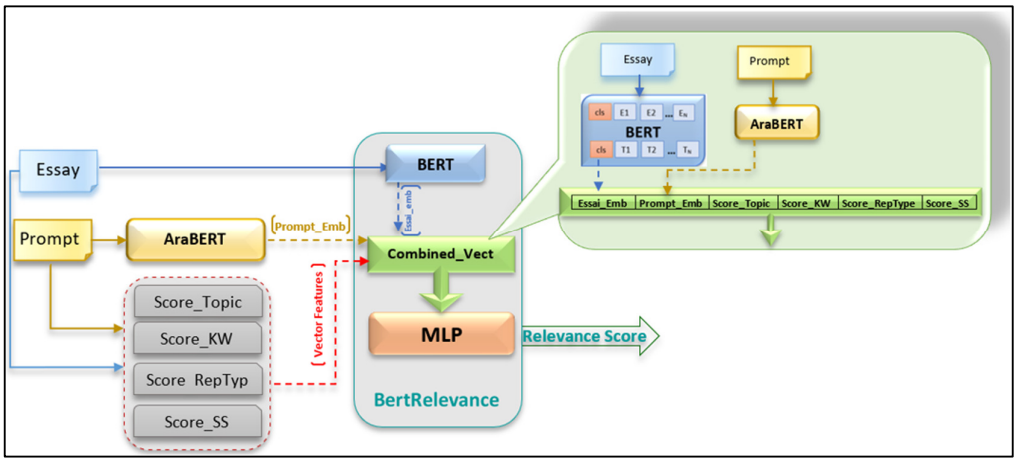

Fig 2. BertRelevance Architecture

Vector_Features

As mentioned in the previous section, the "Combined_Vector" is the concatenation of the contextual representation of the essay wrapped in the token [cls], "Prompt_Emb" representing the prompt and "Vector_Features" covering the handcrafted features. The selection of these features was established based on the recommendations of primary school teachers in Tunisia. We identified three, namely the topic similarity, keywords similarity, and conformity response type. Additionally, we computed the semantic similarity between the essay and the prompt.
We developed a specific model for each criterion, which we will detail below.

a. Topic scoring: the first feature consists of evaluating the conformity of the topic covered in the student's answer to the topic addressed in the prompt. Indeed, the formulation of prompt generally follows a structure in two parts: the context ('السند') and the directive ('التعليمة') (see Figure 1). The first part, 'السند', has the role of introducing the situation and context of issue. Thus, offering the student a better understanding of the overall framework of the prompt. On this basis, we compare the topic introduced in the prompt (in 'السند' more precisely) with the one discussed in the student's essay. To perform this, we employed BERTopic, a topic modeling technique based on the BERT algorithm developed by Grootendorst (2020) [21]. We utilize its pre-trained version on the Arabic corpus proposed by Abuzayed et al (2021) [22]. Based on the theme extracted from both essay and prompt, we calculate a score denoted as "Score_Topic".

b. Keywords scoring: The sharing of some keywords between the essay and the prompt can indicate whether the student's response is in alignment with the question. This is known as the use of the same linguistic lexicon (in Arabic "استخدام نفس المعجم اللغوي"). Using KeyBERT introduced by Grootendorst M (2020) [23], we extracted two lists of keywords, one for the response and the second for the prompt. We then constructed a matrix based on these two lists, with each cell filled by the cosine similarity value between each pair of keywords. A score is subsequently computed from this matrix and is denoted as "Score_KW". See Figure 3.

Fig. 3. Keywords scoring matrix

c. Essay type scoring: writing production exams for 5th and 6th-grade of primary school in Tunisia cover three types of texts: narrative, descriptive and dialogue. As previously mentioned, the prompt statement is divided into two parts: the context ('**السند**') and the directive ('**التعليمة**'). In the directive, the type of text that the student must write is specified (see Figure 1). We have thus developed an algorithm which assigns a score, called "Score_RepType". This score is granted a value of 1 if the student's response respects the type required in the prompt, and 0 otherwise. In fact, a narrative text is composed of verbal phrases. Where the number of verbs, especially action verbs, should be predominant. Additionally, a narrative text is marked by the presence of at least one disruptive element. The pseudo-code of the function that tests whether an essay is narrative or not is illustrated in figure 4.

Algorithm: test_narrative

1.Input: Essay
2.Output: 1 **If** Essay is a Narrative text, 0 **otherwise**
3.Initialize: list_verbs_action = ['يسرع ', 'يرافق ', 'يجلب ', 'يصفق ', ' ،يمسك ','يطرق','يزيد ','يبتلع ','يتقدم', …] ,list_verbs_essay= [] #list of verbs in the essay
, L_V_A= [] #list of action verbs in the essay , DE = 0 #number of disruptive elements , EssayLen =0 #number of tokens in the essay
4. #Tokenize the Essay and update the 'EssayLen' variable

Detecting Verbs of Action

```
4.  #put all verbs into list_verbs_essay
5.  For i in list_verbs_essay:                # Grouping action verbs of the essay in L_V_A
6.    If i in list_verbs_action:
7.      L_V_A.append (i)
```

Detecting the presence of a disruptive element

```
9.  pattern1= (r'لسوء الحظ | لم يدم ذلك طويلا |انقلب الوضع خطرت ببالي فكرة | إسرعان ما إلكن |غير ان  |بيد ان | يتحدث ما لم يكن في الحسبان | …')
10. pattern2= (r" .*(على تلك الحال إذ).*|.*(بعد).*| .*(حتى).*|.*(ماهي الا ساعات).*|.*(حتى).*|.*(ماهي الا لحظات).*|.*(حتى).*|.*(ماهي الا برهة).*|.*(ان).*|.*(ما).*")
11. match1 = re.findall (pattern1, Essay)
12. match2= re.findall (pattern2, Essay)
13. DE= len (match1) + len(match2)
14. nb_verb= len (list_verbs_essay)        #number of total verbs in the essay
15. nb_verb_A = len (L_V_A)                #number of action verbs in the essay
16. if ((nb_verb >= EssayLen * 0.15 and DE >=1) or (nb_verb_A >= nb_verb * 0.13 and  DE>=1)
      or (nb_verb >= EssayLen * 0.15 and nb_verb_A >= nb_verb * 0.13)):
17.   return (1)
18. else:
19.     return (0)
```

Fig. 4. Pseudo-code for the 'test_narrative' function

On the other hand, to categorize a text as a dialogue, it must contain at least one dialogue verb, an expression of opinion and two or more dialogue punctuations. Figure 5 clarifies the pseudo-code of the test dialogue function. A descriptive text is characterized by the presence of several adjectives, pronouns, adverbs of time, place, and manner.

Algorithm: test_dialogue

1.Input: Essay
2.Output: 1 If Essay is a Dialogue or Opinion text, 0 otherwise
3.Initialize: list_ponctuation = ['-','<','>',':','>>','<<', '?', '!', '...'] , list_verbs_dialogue = ['أكد', 'تابع', 'ذكر', 'قال', 'رد', 'استأنف', 'تساءل' ...] , dialogue_mark = 0 , nb_verb = 0, nb_verb_dialogue = 0 , nb_exp_opinion = 0 , list_verbs=[]

calculate the number of punctuation marks

4. For each token in Essay:
5. If token in list_ponctuation:
6. Increment dialogue_mark

calculate the number verb dialogue

7. For each token in Essay:
8. If token is verb:
9. list_verbs.append (token)
10. For each verb in list_verbs:
11. If verb in list_verbs_dialogue:
12. Increment nb_verb_dialogue

calculate the number of opinion expression

13. pattern1= (r'بِالنِّسْبَةِ لِي' , "أَخْتَلِفُ مَعَ' , 'أَعْتَقِدُ أَنَّ' , 'أَظُنُّ أَنَّ' , 'مِنْ وِجْهَةِ نَظَرِي' ,'هَلْ تَعْتَقِدُ أَنَّ', 'مَا رَأْيُكَ فِي')
14. pattern2= (r" .*(أعتقد).*|.*(أوافق).*|(يقنعني).*(الكلام).*|(مَعَ).*(أَخْتَلِفُ).*|(كلامك).*(أوافق).*")
15. match1 = re.findall(pattern1, essai)
16. match2= re.findall(pattern2, essai)
17. nb_exp_opinion = len(match1)+len(match2)
18. **If** (nb_verb_dialogue > 1 or dialogue_mark > 2 or nb_exp_opinion > 1):
19. return 1
20. **else**:
21. return 0

Fig 5. Pseudo-code for the 'test_dialogue ' function

Semantic similarity scoring: in order to improve reliability, we have chosen to calculate the semantic similarity between the essay and the prompt. This decision is backed by the review conducted by Ramnarain-Seetohul et al (2022) [24] as well as the comparative study led by Odunta et al (2018) [36]. We opted for cosine similarity, and we will refer to this feature as Score_Sim".

Extensible version: Our system is extensible (Figure 6), which means we can incorporate an additional model response or suggested correction given by teachers. This additional response, customized to provide an ideal response to the prompt, has the potential to further enhance the results.

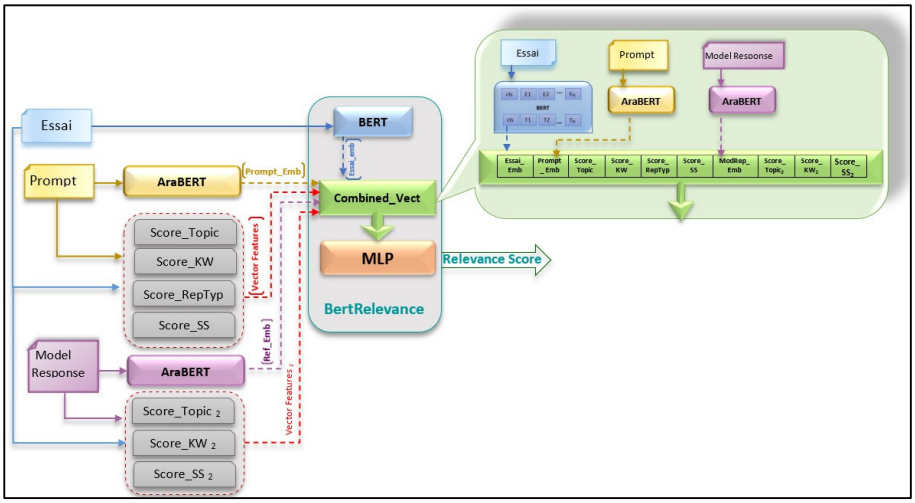

Fig. 6. Extended BertRelevance architecture

We consequently add features modeling the relationship between the essay and the model response, which we identify with index '2'. As a result, the structure of the combined vector will be as represented in Figure 6.

Where:
ModRep_Emb: the embedding vector representing the model response.
Score_Topic$_2$: the score determined by evaluating the conformity of topics between the essay and the model response.
Score_KW$_2$: the score granted for keyword similarity between the essay and the model response.
Score_SS$_2$: the cosine similarity between the essay and the model response.

Experiments and evaluation

Dataset

The problem of insufficient data is a major concern for many researchers in Arabic NLP. In some cases, this leads to the unavailability of open-source dataset, while in other cases, data quality is unsatisfactory.

We have also encountered this problem, with the absence of an open-source dataset for long essays in the Arabic language. Consequently, we set out the task of constructing our own dataset. To do so, we started by limiting our target in order to guarantee a unified and adapted evaluation grid for all copies. We chose to consider the responses of students in the 5th and 6th years of primary education. Because these levels are almost similar and follow the same evaluation grid defined by the Tunisian Ministry of Education (figure 7).

الأعداد المسندة Assigned scores					الجملة Total	مؤشرات المعيار Standards Norm	المعيار Norms
أ: جيد جدا A : Very good	ب: جيد B : Good	ج: مقبول C : Acceptable	و: دون الأدنى D : Below Minimum	ه: غير مقبول F : Unacceptable			
2	1.5	1	0.5	0	4	توافق الإنتاج مع التعليمة Compliance of the essay with the directive	الملاءمة Relevance
2	1.5	1	0.5	0		توافق الإنتاج مع Compliance of the essay with the context	
1	0.75	0.5	0.25	0	5	استعمال الروابط استعمالا سليما Use links properly	سلامة بناء النص Integrity of text structure
1	0.75	0.5	0.25	0		ترتيب الأحداث Order of events	
1	0.75	0.5	0.25	0		احترام قواعد الرسم Respect writing rules	
1	0.75	0.5	0.25	0		اكتمال البنية السردية Complete narrative structure	
1	0.75	0.5	0.25	0		استعمال الأبنية اللغوية استعمالا سليما Using linguistic structures correctly	
1	0.75	0.5	0.25	0	3	الإغناء بالوصف Enrichment by description	التصرف في نمط الكتابة Mastery of essay style
1	0.75	0.5	0.25	0		الإغناء بالحوار Enrichment by dialogue	
1	0.75	0.5	0.25	0		إحداث مفارقة سردية Creating a narrative paradox	
1	0.75	0.5	0.25	0	4	استعمال معجم فصيح Use an eloquent dictionary	ثراء اللغة والطرافة Language Richness and Wit
1	0.75	0.5	0.25	0		ظهور فكرة متميزة أو أكثر The emergence of one or more distinct ideas	
1	0.75	0.5	0.25	0		إستعمال تراكيب متنوعة Use various compositions	
1	0.75	0.5	0.25	0		تصرف طريف في حبكة النص A funny behavior in the plot of the text	
2	1.5	1	0.5	0	4	وضوح الكتابة Clarity of writing	حسن العرض Effective presentation
1	0.75	0.5	0.25	0		سلامة التنقيط Punctuation accuracy	
1	0.75	0.5	0.25	0		تمايز الفقرات Distinctiveness of paragraphs	

Fig 7. Scoring scheme for 5th and 6th grades

We collected 260 essays on different topics. The length of the essays varied between 366 and 19 words. All copies were retyped on the computer exactly as they were handwritten, including the same errors. Each essay was evaluated by a primary school teacher based on the criteria in Figure 7. However, upon analyzing the collected data, we observed that our data presented two downsides:

limited size and an imbalance. To solve these problems, we use the "Random Over Sampling" (RO) method as implemented by Branco (2022) in the 'Imbalanced Learning Regression' Python package. We were finally able to increase our dataset from 260 unbalanced essays to 380 balanced essays as shown in Figure 8.

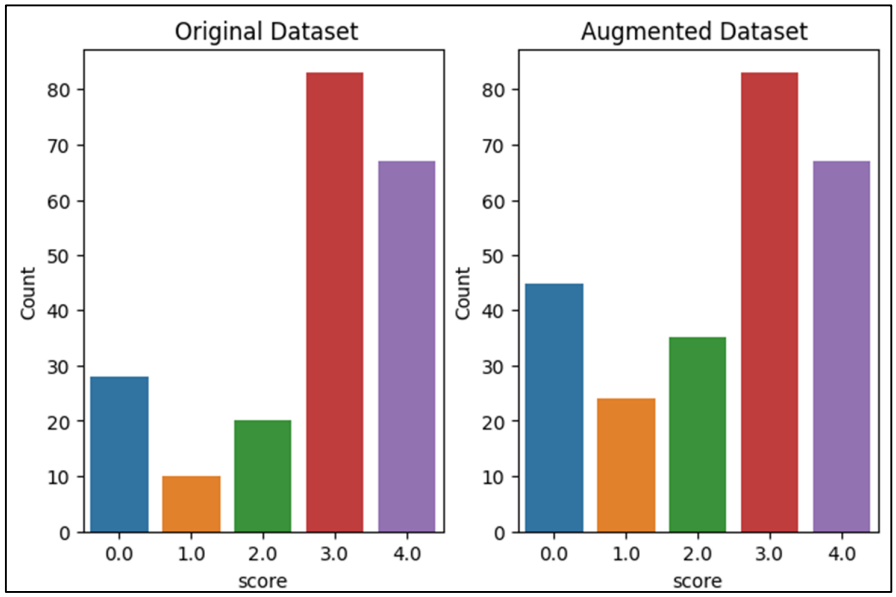

Fig 8. Dataset distribution

Experimental setup

In this section, we present our multiple experiments. We conducted tests on different versions of our model to study the impact of adding features on the relevance score prediction results. We maintain the basic structure of our model and vary the inputs, consequently modifying the "Combined_Vector" each time, as illustrated in Table 1.

Table 1. Combined_Vector variations

Test N°	Input	Vector combined
1	Essay + Prompt	[Essay_emb, Prompt_Emb]
2	Essay+ Prompt + Prompt_Features	[Essay_emb, Prompt_Emb, Score_Topic, Score_KW, Score_RepTyp, Score_Sim]
3	Essay + Prompt_Features	[Essay_emb, Score_Topic, Score_KW, Score_RepTyp, Score_Sim]
4	Essay + Prompt + Prompt_Features + ModRep+ ModRep_Features	[Essay_emb, Prompt_Emb, Score_Topic, Score_KW, Score_RepTyp, Score_Sim, RepMod_Emb, Score_Topic2, Score_KW2, Score_SS2]
5	Essay + Prompt_Features + ModRep + ModRep_Features	[Essay_emb, Score_Topic, Score_KW, Score_RepTyp, Score_Sim, RepMod_Emb, Score_Topic2, Score_KW2, Score_SS2]
6	Essay + Prompt_Features + ModRep_Features	[Essay_emb, Score_Topic, Score_KW, Score_RepTyp, Score_Sim, Score_Topic2, Score_KW2, Score_SS2]

**Prompt_Features: features extracted from essay and prompt, ModRep: model response, ModRep_Features: features extracted from essay and model response

We proceeded to test our model in two main stages: first by inputting the prompt with the essay (Test N° 1,2 and 3) then by integrating the model response in addition (Test N° 4,5 and 6).

In test N°1, we introduced the prompt along with the essay (Prompt_Emb). Then, in Test N°2, we integrated in addition the various features extracted from the essay and prompt (Prompt_Features). Furthermore, in Test N°3, we

excluded the prompt and kept the essay with the features representing the relationship between the essay and the prompt (Prompt_Features).

In Test N°4, we extended our approach by adding to the inputs of Test N°2 a model response (ModRep_Emb) as well as the features extracted from the essay and this response (ModRep_Features). We then excluded the prompt in Test N°5. Finally, in

Test N°6, we only associated the features linked to the prompt (Prompt_Features) and those linked to the model response (ModRep_Features).

Scoring metrics

To evaluate the results of the various experiments carried out, we used the Pearson's correlation metric noted 'r' (1). As highlighted by Plevris et al (2022), this is the most commonly used metric in the context of regression, as it measures the strength of the relationship between two variables. The Pearson correlation assigns a value of 1 in the case of a positive correlation, and -1 in the case of a negative correlation. The closer the value is to zero, the more independent the variables are. In our case, this measure tells us the degree to which the grade predicted by our system matches the grade awarded by the teachers.

$$r = \frac{\sum (\alpha - \tilde{\alpha})(y - \bar{y})}{\sqrt{\sum (\alpha - \tilde{\alpha})^2 \sum (y - \bar{y})^2}} \quad (1)$$

Hyperparameters

We experimented with various hyperparameters' settings to optimize our results. After evaluating multiple combinations of these parameters, we identified the most effective settings which are summarized in table 2.

Table 2. Selected values

Test N°	**1, 2, 3**	**4, 5, 6**
Embedding	AraBert	
Loss function	PRelu	
Activation function	MSE	
Batch size	1	
Epochs	50	
Learning rate	0.0001	0.001
Optimizer	Adamax (lr=2^e-3, eps=2^e-5)	Adamax (lr=3^e-3, eps=2^e-5)

Experimental results

We present in the following table the results of various experiments conducted:

Table 3. Experimental results

Test N°	**Inputs**	**Correlation**
1	Essay + Prompt	0.48
2	Essay+ Prompt + Prompt_Features	0.73
3	Essay + Prompt_Features	**0.77**
4	Essay + Prompt + Prompt_Features + ModRep + ModRep _Features	**0.85**

5	Essay + ModRep + Prompt_Features + ModRep_Features	**0.88**
6	Essay + Prompt_Features + ModRep_Features	0.80

To evaluate the performance of our "BertRelevance" model, we conducted an initial series of tests. We began by incorporating both the prompt and the essay, which resulted in a correlation of 0.46. Then the addition of features extracted from both inputs significantly improved the results, reaching a correlation of 0.73. In the subsequent test, by removing the prompt as an input, we achieved a correlation of 0.77.

Based on this series of tests, we can observe that including the features enhances BERT's ability to predict the appropriate score. However, for better orientation, it is more interesting to provide only the features essential to the task.

We continue with a second series of experiments in which we enrich our inputs with a model response. Initially, we added to the triplet (essay, prompt, and features) the model response (ModRep) along with the features (ModRep_Features) extracted from both the essay and this response. This resulted in a significant improvement, with the correlation rising to 0.85. In the next test, we removed the prompt, but kept the features extracted from it. In this configuration, we obtained our best result with a correlation of 0.88. Finally, we explore the possibility of removing both the prompt and the model response, while retaining only the features (Prompt_Features and ModRep_Features) associated with the essay. In this case, the correlation was 0.80.

According to the results of the second series of tests, they assure us that the choice of adding features must be made judiciously.

From our analysis of our various experiments, it's clear that including handcrafted features significantly improves BERT's performance. This can be explained by the fact that these features represent additional pertinent information for the relevance scoring which boosts the model and improves the correlation.

However, it is important to note that the selection of features to be integrated should be made prudently. At times, the incorporation of inappropriate features can lead to confusions in BERT's learning process. In our case, the model response embedding is beneficial as it provides additional information useful for relevance detection. While the prompt embedding may not be as informative as this information is already present in the context of the essay. Here, the prompt embedding is redundant information which is already implicitly captured in the context. This duplication of information can disrupt model training, leading to a decrease in correlation.

Conclusion

In this paper, we have presented a new approach for scoring the relevance of the essay content regarding the prompt, in the context of automated essay scoring system. Our proposal deals with essays in the form of long Arabic text and is based on the deep learning algorithm BERT (specifically AraBERT), enhanced by handcrafted features. Our proposal has achieved promising results, with a correlation of 0.88 with human score. The various tests we carried out show that adding features can considerably improve results. This can be explained by the fact that the Arabic version of BERT is pre-trained on a considerably limited corpus. As a result, these carefully selected features play a strengthening role. Therefore, if we assume that BERT's pre-training corpus is sufficiently large, this type of features helps to guide BERT in its training. Finally, based on our experience, we conclude that the use of BERT, enhanced with handcrafted features for scoring the relevance of Arabic essays to the prompt, presents encouraging prospects. However, it is essential to highlight the importance of making a careful choice when selecting handcrafted features.

In light of these promising results, we intend to continue improving our performance by exploring potential new features and expanding our training dataset. In the long term, our aim is to integrate this work into a larger project dedicated to the automatic scoring of Arabic language essays.

References

- Abdul-Mageed, M., Elmadany, A., & Nagoudi, E. M. B. (2020). ARBERT & MARBERT: deep bidirectional transformers for Arabic. arXiv preprint arXiv:2101.01785.
- Abuzayed, A., & Al-Khalifa, H. (2021). BERT for Arabic topic modeling: An experimental

study on BERTopic technique. Procedia computer science, 189, 191-194.

- Alobed, M., Altrad, A. M., & Bakar, Z. B. A. (2021, June). An Adaptive Automated Arabic Essay Scoring Model Using the Semantic of Arabic WordNet. In 2021 2nd International Conference on Smart Computing and Electronic Enterprise (ICSCEE) (pp. 45-54). IEEE.
- Alqahtani, A., & Alsaif, A. (2019, December). Automatic evaluation for Arabic essays: a rule-based system. In 2019 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT) (pp. 1-7). IEEE.
- Alqahtani, A., & Al-Saif, A. (2020, December). Automated Arabic Essay Evaluation. In Proceedings of the 17th International Conference on Natural Language Processing (ICON) (pp. 181-190).
- Antoun, W., Baly, F., & Hajj, H. (2020). Arabert: Transformer-based model for arabic language understanding. arXiv preprint arXiv:2003.00104.
- Azmi, A. M., Al-Jouie, M. F., & Hussain, M. (2019). AAEE–Automated evaluation of students' essays in Arabic language. Information Processing & Management, 56(5), 1736-1752.
- Beseiso, M., Alzubi, O. A., & Rashaideh, H. (2021). A novel automated essay scoring approach for reliable higher educational assessments. Journal of Computing in Higher Education, 33, 727-746.
- Branco, P. (2022). ImbalancedLearningRegression-A Python Package to Tackle the Imbalanced Regression Problem.
- Crossley, S. A., Varner, L. K., Roscoe, R. D., & McNamara, D. S. (2013). Using automated indices of cohesion to evaluate an intelligent tutoring system and an automated writing evaluation system. In Artificial Intelligence in Education: 16th International Conference, AIED 2013, Memphis, TN, USA, July 9-13, 2013. Proceedings 16 (pp. 269-278). Springer Berlin Heidelberg.
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
- Farag, Y., Yannakoudakis, H., & Briscoe, T. (2018). Neural automated essay scoring and coherence modeling for adversarially crafted input. arXiv preprint arXiv:1804.06898.
- Grootendorst, M. (2020). BERTopic: Leveraging BERT and c-TF-IDF to create easily interpretable topics, Zenodo.
- Gu, K., & Budhkar, A. (2021, June). A package for learning on tabular and text data with transformers. In Proceedings of the Third Workshop on Multimodal Artificial Intelligence (pp. 69-73).
- Inoue, G., Alhafni, B., Baimukan, N., Bouamor, H., & Habash, N. (2021). The interplay of variant, size, and task type in Arabic pre-trained language models. arXiv preprint arXiv:2103.06678.
- Jong, Y. J., Kim, Y. J., & Ri, O. C. (2023). Review of feedback in Automated Essay Scoring. arXiv preprint arXiv:2307.05553.
- Li, X., Chen, M., Nie, J., Liu, Z., Feng, Z., & Cai, Y. (2018). Coherence-based automated essay scoring using self-attention. In Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data: 17th China National Conference, CCL 2018, and 6th International Symposium, NLP-NABD 2018, Changsha, China, October 19–21, 2018, Proceedings 17 (pp. 386-397). Springer International Publishing.
- Machhout, R. A., Zribi, C. B. O., & Bouzid, S. M. (2021, December). Arabic Automatic Essay Scoring Systems: An Overview Study. In International Conference on Intelligent Systems Design and Applications (pp. 1164-1176). Cham: Springer International Publishing.
- Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. Research Methods in Applied Linguistics, 2(2), 100050.
- Oduntan, O. E., Adeyanju, I. A., Falohun, A. S., & Obe, O. O. (2018). A comparative analysis of euclidean distance and cosine similarity measure for automated essay-type grading. Journal of Engineering and Applied Sciences, 13(11), 4198-4204.
- Palma, D., & Atkinson, J. (2018). Coherence-based automatic essay assessment. IEEE Intelligent Systems, 33(5), 26-36.
- Persing, I., & Ng, V. (2014, June). Modeling prompt adherence in student essays. In Proceedings of the 52nd Annual Meeting of the Association for Computational

Linguistics (Volume 1: Long Papers) (pp. 1534-1543).

- Plevris, V., Solorzano, G., Bakas, N. P., & Ben Seghier, M. E. A. (2022, November). Investigation of performance metrics in regression analysis and machine learning-based prediction models. In 8th European Congress on Computational Methods in Applied Sciences and Engineering (ECCOMAS Congress 2022). European Community on Computational Methods in Applied Sciences.
- Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring systems: a systematic literature review. Artificial Intelligence Review, 55(3), 2495-2527.
- Ramesh, D., & Sanampudi, S. K. (2022, November). Coherence Based Automatic Essay Scoring Using Sentence Embedding and Recurrent Neural Networks. In International Conference on Speech and Computer (pp. 139-154). Cham: Springer International Publishing.
- Ramnarain-Seetohul, V., Bassoo, V., & Rosunally, Y. (2022). Similarity measures in automated essay scoring systems: A ten-year review. Education and Information Technologies, 27(4), 5573-5604.
- Safaya, A., Abdullatif, M., & Yuret, D. (2020). Kuisail at semeval-2020 task 12: Bert-cnn for offensive speech identification in social media. arXiv preprint arXiv:2007.13184.
- Salim, Y., Stevanus, V., Barlian, E., Sari, A. C., & Suhartono, D. (2019, December). Automated English digital essay grader using machine learning. In 2019 IEEE International Conference on Engineering, Technology and Education (TALE) (pp. 1-6). IEEE.
- Tay, Y., Phan, M., Tuan, L. A., & Hui, S. C. (2018, April). Skipflow: Incorporating neural coherence features for end-to-end automatic text scoring. In Proceedings of the AAAI conference on artificial intelligence (Vol. 32, No. 1).
- Zupanc, K., & Bosnic, Z. (2014, December). Automated essay evaluation augmented with semantic coherence measures. In 2014 IEEE International Conference on Data Mining (pp. 1133-1138). IEEE.