Automated Arabic Essay Evaluation

Abeer Alqahtani
Department of Computer Science
Al Immam Mohammad Ibn Saud
Islamic University(IMSIU)
Riyadh, Saudi Arabia
aakalqahtani@sm.imamu.edu.sa

Amal Alsaif
Department of Computer Science
Al Immam Mohammad Ibn Saud
Islamic University(IMSIU)
Riyadh, Saudi Arabia
asmalsaif@imamu.edu.sa

Abstract

Although the manual evaluation of essays is a time-consuming process, writing essays has a significant role in assessing learning outcomes. Therefore, automated essay evaluation represents a solution, especially for schools, universities, and testing companies. Moreover, the existence of such systems overcomes some factors that influence manual evaluation such as the evaluator's mental state, the disparity between evaluators, and others. In this paper, we propose an Arabic essay evaluation system based on a support vector regression (SVR) model along with a wide range of features including morphological, syntactic, semantic, and discourse features. The system evaluates essays according to five criteria: spelling, essay structure, coherence level, style, and punctuation marks, without the need for domain-representative essays (a model essay). A specific model is developed for each criterion; thus, the overall evaluation of the essay is a combination of the previous criteria results. We develop our dataset based on essays written by university students and journalists whose native language is Arabic. The dataset is then evaluated by experts. The experimental results show that 96% of our dataset is correctly evaluated in the overall score and the correlation between the system and the experts' evaluation is 0.87. Additionally, the system shows variant results in evaluating criteria separately.

1 Introduction

Automated essay scoring (AES) or so-called automated essay evaluation (AEE) systems came to facilitate the evaluation task of students' writings. Assigning questions that demonstrate writing skills, such as linguistic skills and creativity, is crucial. However, the evaluation process is laborious, particularly with a large number of essays, such as during examination boards.

AES systems assist essay authors, editorial boards, publishers, and newspaper editors by overcoming some of the shortcomings of traditional evaluation. For instance, AES systems reduce variabilities between instructors' viewpoints or biases resulting from a good point in essay that causes an evaluator to ignore other mistakes (Janda et al., 2019). Moreover, AES systems are tools that can assist both new instructors and students for training and improving writing skills.

For the English language, there have been many studies drafted on AES in addition to the development of commercial applications used in English-learning institutes. In comparison, AES systems for Arabic seem restricted to short-answer questions with predefined answer models from instructors. This limitation stems from complexities of the Arabic language and a lack of Arabic natural language processing (NLP) resources.

In this paper, we attempt to fill this gap by providing an Arabic essay evaluation system using a machine learning method that does not require a model essay.

2 Related Work

In English, there are several related efforts. For example, the project essay grader (PEG) is one of the earliest scoring systems. The system was based on a statistical method to predict the score. PEG

Proceedings of the 17th International Conference on Natural Language Processing, pages 181–190
Patna, India, December 18 - 21, 2020. ©2020 NLP Association of India (NLPAI)

succeeded in predicting the surface structure of an essay, but it did not meet the semantic criteria (Chung & O'Neil, 1997; Kukich, 2000).

In the intelligent essay assessor (IEA), latent semantic analysis (LSA) is used to evaluate essay content. In addition, IEA can be used to evaluate writing style and detect plagiarism (Dessus and Lemaire, 2001). The E-rater (Burstein, 2003) and IntelliMetric (Elliot, 2003) are evaluation systems that rely on linguistic features extracted by NLP techniques to evaluate common criteria.

In contrast to previous systems, the Bayesian essay test scoring system (BETSY) is non-commercial and can be used for research. BETSY uses Bayesian models to evaluate essays for content and style (Dikli, 2006).

Although commercial systems have restrictions to accessing system details, AES systems have attracted the attention of many researchers. Most studies, such as Surya et al. (2018), have focused on the automatic student assessment prize (ASAP) dataset.

Surya et al. (2018) used machine learning with nine surfaces, deep features, and three algorithms including support vector machine (SVM), k-nearest neighbors (kNN), and linear regression. The obtained accuracy ranges from 73% to 93% according to the dataset class and the algorithms used. In addition, Alikaniotis, Yannakoudakis, and Rei (2016) and Dong and Zhang (2016) have employed deep-learning algorithms and obtained encouraging results.

In Arabic, there are several studies for short-answer questions scoring. Nahar & Alsmadi (2009) based on light stemming and assigning weights over words in the model answer. For the system to consider semantics, the instructor must manually attach the synonyms. In comparison, Gomaa and Fahmy (2014) combined similarity measures, such as string, corpus, and knowledge-based similarity for 610 Arabic short answers, which were translated to English due to the lack of Arabic resources. However, this approach requires great effort for translation and then scoring.

Apart from semantics, Al-Shalabi (2016) proposed a scoring system for online exams using stemming and Levenshtein string-based similarity measures. Shehab, Faroun, and Rashad (2018) conducted a comparison between several string-based (Damerau–Levenshtein and N-gram) and corpus-based similarity measures (LSA and DISCO) on 210 answers. They found that applying N-gram with removing stop words produced the best results.

Other than short-answer scoring, we only came across three studies for essay scoring (Alghamdi et al., 2014; Azmi, Al-Jouie, & Hussain, 2019; Alqahtani & Alsaif, 2019).

First, Alghamdi et al. (2014) conducted a study based on a linear regression algorithm. They used LSA, the number of words, and spelling mistakes to predict an essay's score. The proposed system was applied to 579 essays collected from undergraduate university students and scored by two instructors. The results showed that 96.72% of essays were scored correctly, while the correlation between the system and manual scoring was 0.78, which is close to the value of 0.7 obtained by inter-human correlation.

Second, Azmi, Al-Jouie, and Hussain (2019) collected 360 essays from students in middle and high schools. The essays were evaluated by two school instructors according to criteria obtained from a questionnaire given to instructors. The criteria can be presented as semantic analysis, writing style, and spelling mistakes. However, there was a difference in assigning each criterion weight, but the majority was 5%, 40%, and 10% for semantic analysis, writing style and spelling mistakes, respectively. The proposed system was based on LSA to evaluate semantics while rhetorical structure theory (RST) and other features were used to evaluate the writing style. Finally, the system employed AraComLex (Attia, Pecina, Toral, Tounsi, & Genabith, 2011) to detect spelling mistakes. The system achieved an accuracy of 90% while the correlation to the manual evaluation was 0.756, which outperformed the inter-human correlation of 0.709.

Both Alghamdi et al. (2014) and Azmi, Al-Jouie, and Hussain (2019) relied on the LSA approach, which requires domain-representative essays. Furthermore, outputs of Alghamdi et al. (2014) are given as an overall score without details about the score for each criterion. In contrast, Azmi, Al-Jouie, and Hussain (2019) employ rules to evaluate writing style and spelling mistakes separately. However, half of the assigned score is based on LSA, which in turn requires pre-defined models.

Recently, (Alqahtani & Alsaif, 2019) proposed a rule-based system to evaluate Arabic essays without the need to train on domain-representative essays. They adopted an evaluation criteria scheme based on Arabic literary resources and university instructors' experiences as following: spelling, grammar, structure, cohesion, style and punctuation marks. Then, scheme was given to two experts to evaluate 100 essays for university students. The system follows a set of rules to evaluate the previous criteria based on some facts and analyses to evaluate the overall score beside the specific criteria scores. The system accuracy was 73% in the overall score while there were variations in evaluating criteria separately. Although this study does need a model essay, it is limited to set of rules that does not include semantic.

Therefore, in this paper, we propose an Arabic essay evaluation system does not need to train on predefined essays represent domain by using a machine learning algorithm and a wide range of features to evaluate the specific criteria besides the overall score.

3 Dataset

Our dataset can be classified into three parts: essays written by undergraduate/graduate Arabic native students with university-level; unedited essays written to be published in one of the Saudi newspapers and essays have been published in different newspapers. Any handwritten copies have been retyped by computer exactly as they were written. Altogether, our dataset contains 200 (MSA) essays with an approximately average length of 250 words (3KB) in several topics.

This dataset has been given to two Arabic experts hold master's degree in Arabic language to evaluate essays following the criteria and the evaluation rubric shown precisely in (Alqahtani & Alsaif, 2019) which include: spelling, grammar, structure, coherence and cohesion level, style and punctuation marks.

Therefore, in spelling criterion, evaluators concern about the correct spell of words; therefore, they detect each spelling mistake and classify it to one of four types: mistakes on <*hmzp*/همزة/ء>, replacement letters, extra letters and neglecting letters. For structure criterion, they check the existence of four essential parts; title, introduction, body and conclusion. In coherence, they evaluate the coherence between the title and remaining parts and check the cohesion between essay parts. This criterion also concerns with the correct use of connectives. To evaluate style of essay, they consider, words repetition, length of sentences, word choice and avoiding lengthy speech. In addition, they evaluate the punctuation marks according to the Arabic rules.

4 Automated Essay Evaluation

We intend to model Arabic essay evaluation based on the supervised linear model Support Vector Regression (SVR) (Awad & Khanna, 2015) along with different levels of features. We develop a specific model for each criterion. The overall evaluation of an essay will be a combination of the models' outputs. We follow this procedure to ensure that each criterion model takes the full advantage of the used features hence more accurate in the overall score. In addition, existence of separated model per criterion will assist in the future to provide a valuable feedback for the user. In the preprocessing step, we applied normalization, then stemming using Buckwalter stemmer (Buckwalter, 2004). Our features extracted by considering the criteria followed by humans. According to the system output, which is numeric scores, we try a wide range of features represented by numbers at different levels including:

A. Surface features

The surface features demonstrate only features dealing with the text itself such as word frequency. Each feature is followed by abbreviation for ease of reference as follows:

Essay Length (F1): is measured by number of tokens/words result of white space tokenization of essay.

Number of paragraphs (F3, F4) and sentences (F5): we considered each line as a paragraph so that when the writer moves to the next line because there is no space it is considered as one line (F3). Additionally, we checked if number of paragraphs is greater than one or not (F4). To count number of sentences (F5), we divided essays by period, comma and a list of connectives that usually used to connect two sentences in Arabic according to (Alsaif, 2012) study.

Words per essay parts (F6), (F7), (F8), (F9), and (F16): knowing the length of a specific part of essay, may lead to identifying its role in the essay, (e.g., when the first part is the shortest, that may indicate it is the title). Generally, distributing features over essay parts lead to detect the effective features to evaluate a specific criterion. So, in separated features, we counted the number of words in first paragraph/title (F6), second paragraph /introduction (F7), last paragraph/conclusion (F8), and number of paragraphs in the middle/body (F9). Also, we checked if the first paragraph length is less than or equal to ten words (F16).

Average, maximum, and minimum length (F10–F15): in separated features, based on number of words we calculated the average length of sentences (F10), longest paragraph (F12) and the shortest paragraph (F13). Likewise, we calculated the longest sentence (F14) and the shortest sentence (F15).

Paragraph has a specific mark (F19–F22): for example, the presence of some marks may indicate the essence of paragraph. Separately, we checked each essay part if it contains parentheses, colon or question mark (F19–F22).

Number of <hmzp/همزة/ء> (F22) in essay: we counted words that contain *hmzp on AlOlf /أ* and إ, *hmzp on AlwAw /ؤ*, *hmzp on line/ء* and *hmzp on nbrp/ئ* (F22) to assist in spelling evaluation.

B. Syntactic and Morphological features

Parts of speech (POS) frequency in essay (F2–F99): by using the (POS) tag provided by the MADAMIRA analyzer (Pasha et al., 2014), we counted the amount of punctuation (F23), pronouns (F24), prepositions (F25), verbs (F26), nouns (F27), adjectives (F28), adverbs (F29) and numbers (F30) in the whole essay. In addition, we calculated these features for essay parts (paragraphs and sentences). Also, we checked paragraphs and sentences that start with a specific POS (adjectives and prepositions) separately (F31–F99).

Number of nominal and verbal sentence (F100) and (F101): for each sentence, we checked the first three words. If the sentence included a verb, it was considered as a verbal (F100), otherwise; it was considered as a nominal phrase (F101).

Spelling mistakes in the whole essay (F102–F109): In line with Alqahtani and Alsaif (2019), we used the FARASA spell checker (*FARASA: Advanced Tools for Arabic*, 2019) to classify the mistakes that FARASA provided into four classes: mistakes on *hmzp*, replacement letters, extra letters, and omission letters. Therefore, as features, we counted the number of spelling mistakes in any type of *hmzp* (F102), mistakes in replacement (F103), extra letters (F0104), or omissions (F105). Additionally, considering all the previous features, if its value was more than or equal to one, we assigned 1 to indicate a mistake; otherwise, we assigned 0 (F106–F 109).

Aljzm/الجزم particles (F110) and (F111): we counted number of *aljzm* particles in the essay as they cause a change in the subsequent verbs *(lm/لم, lA AlnAhyp /لا الناهية, lA AlOmr /لام الأمر, lmA/لما)* (F110) as well as the number of times they were followed by a verb (F200). Also, we counted the number of cases of *aljzm* particles followed by a plural verb ending with *n/ن*, as this case affects the word form (F201) (Ali, 2019).

kAn wOxwAthA /Kana and her Sisters/كان وأخواتها(F112) and (F113): we counted the number of *kAn wOxwAthA*, which include: (*kAn/كان, ODHY /أضحى, mAzAl/مازال, lys/ليس, mAZl/ماظل, OmsY /أمسى, mAft}/مافتئ, bAt/بات, SAr/صار, Zl/ظل, mAAnfk/ماانفك, mAbrH/مابرح, mAdAm/مادام, OSbH أصبح*) as they may affect the surrounding word forms (F112) (Ali, 2018).

In~ wOxwAthA /Inna and her Sisters/إن وأخواتها (F114): Moreover, we counted number of *In~ wOxwAthA*: (On~ /أن, In~/إن, kOn /كأن, lkn/لكن, lyt/ليت, lA/لا, lEl/لعل) in the essay for their effect on words forms (Ali, 2018).

Morphological mistakes (F117–F119): to count the number of words that were morphologically incorrect, we counted the number of words that could be analyzed by MADAMIRA (F117), the number of words that could not be analyzed by MADAMIRA (F118) and the number of words that their lemmas were not included in alWaseet or Contemporary dictionaries using SAFAR platform (*SAFAR: Software Architecture For ARabic*, 2013). In addition, we checked the style of plural words so, if the type of word was (p/plural) or the word ended with (ات) and its lemma ended with (ة/ه), but it does not belong to alWaseet or Contemporary dictionaries, then the number of mistakes in sound feminine plural increases (F119). These features may assist in evaluate spelling and style criteria.

C. Lexical features

Number of words without stop words (F123): refers to the number of words without stop words frequently used in Arabic text.

Introduction and conclusion keywords (F124–F127): usually, the introductory section may include some keywords used to pave a topic. Likewise, the writer may use specific words to conclude or summarize the essay. Therefore, we have two features, the first for checking if the first or second paragraph contains introductory keywords such as (*bdAyp/بداية*, *ntHdv/نتحدث*, *ntklm/نتكلم*, *nstErD/نستعرض*, or *AlmwDwE/الموضوع*, etc.) (F124). We did the same by checking the last paragraph for the conclusion (F125). The common words used to conclude in our dataset and Arabic essays generally include *(OrY/أرى, OxyrA/أخيرا, Orjw/أرجو, wjhp nZr/ وجهة نظر, OqtrH/أقترح, OtmnY/أتمنى, etc.)* with their derivations. In (F126–F127). Furthermore, we checked for the existence of inappropriate words wrongly used to start or conclude an Arabic essay such as *(bsm Allh AlrHmn AlrHym/ بسم الله الرحمن الرحيم, OmA bEd/ أمابعد, AlHmdllh rb AlEAlmyn/ الحمدلله رب العالمين, SlY Allh wbArk/صلى الله وبارك, etc.)* which usually used in other types of writing in Arabic. ***Arabic Lexicon Features (F128–F133):*** we relied on four Arabic lexicons to extract (F128–F133): *The Contemporary Arabic Language Dictionary*; alWaseet lexicon; the Arabic Wordlist for spellchecking (Attia, Pecina, & Samih, 2012) which contains 9 million words automatically generated from the AraComLex open-source finite-state transducer (30,000 lemmas), and a billion-word corpus; and Obsolete Arabic Words (Attia, Pecina, Toral, Tounsi, & Genabith, 2011) which includes obsolete words or words that are not in contemporary use, in the Buckwalter Morphological Analyzer database. As separate features, we checked the number of words that belong to each of the obsolete list, alWaseet, and Contemporary lexicons, as well as words that do not belong to the spellchecking list. We proposed these features to evaluate spelling and style criteria.

Punctuation features (F134–F161): for each punctuation mark, we counted its frequency in the essay. So, in separate features, we counted the frequency of each of the following: question mark (F134), exclamation (F135), period (F136), comma (F137), semicolon (F138), quotation mark (F139), parentheses (F140), dash (F141) and colon (F142). Also, we counted the number of times the writer repeats the same punctuation mark in one use (F142) such as a repeating period (....) or question mark (???), which represents one of the common mistakes in Arabic writing. Furthermore, we have additional features related to each punctuation mark that can be broken down into three categories: correct use, missing use, and incorrect use of a punctuation mark. For the correct use of a question mark (F143), we counted the number of times a sentence contains question tools, including *hl/هل*, *kyf/كيف*, *mA*A/ماذا*, *lmA*A/لماذا*, *mA/لم*, *km/ كم*, *mtY/متى*, or *Ayn/أين*, along with a question mark. The question mark was considered missing if a sentence contained one of the question tools yet was missing a question mark (F144). In case of an incorrect usage, we counted the number of times a sentence contained a question mark without the existence of a question tool (F145).

For correct use of the exclamation mark, we counted the number of times an exclamation existed in a sentence containing one of the exaggerating styles, such as *yAlyt/ياليت*, *b}s/بئس*, *rA}E/رائع*, or *llh dr~/لله در*, or contained a word in the pattern *mA OfEl/ما أفعل* (F146). A missing use was considered if one of these keywords existed while the exclamation mark was absent (F147). For an incorrect use, we counted the number of times an exclamation existed while the previous indicators were missing (F148). For semicolons, to detect correct usage we checked the word following the semicolon. If the word had a causative meaning, such as *lOn/لأن*, *bsbb/بسبب*,*ky/لكي*, or the word started with the clitic *l/لـ* or *f/فـ*, then the number of correct uses of the semicolon increases (F149). A missing use considered when the previous indicators existed and the semicolon mark was missing (F150). For the wrong use of the semicolon, we counted the number of times a semicolon was not followed by those causative indicators (F151). Also, we considered comma as an incorrect use if it was involved in the paragraph containing discourse connectives as presented in Alsaif (2012) (F152). A comma also was classified as misused if a paragraph containing connectives was missing a comma (F153), and an incorrect use was considered in the case of a comma followed by causative indicators.

For the period mark, we counted the number of correct usages if each paragraph ended with a period (F154), and we increased the number of missing period marks if the paragraph did not end with a period. An incorrect usage of the period was considered when a paragraph contained a period before the end (F155). Moreover, we counted the number of correct uses of the colon by checking the existence of some words such as *mvAl/مثال*, *Al|typ/التالية*, *AltAlyp/الآتية*, or *mAyly/مايلي* with a colon mark. Also, a colon was considered correct if it was involved in a sentence containing a word referring to reported speech based on list of attribution cues, as seen in a study by (Alsaif et al. 2018) (F156). We counted the number of times a colon was missing when indicators were present yet a colon was absent (F157). Conversely, the use of a colon was considered wrong if a colon mark was present while indicators were missing (F158). Finally, for quotation marks, we counted the number of correct uses by checking cases in which quotation marks were preceded by one of the words referring to reported speech, using the list of attribution cues and where opening and closing pairs of quotations were placed (F159). A missing use of a quotation mark was considered when at least one word of the list mentioned was present yet quotation mark pairs or a single one was missing (F160). An incorrect use was considered when a quotation mark was present without a cue (F161).

D. Semantic features

Unlike traditional dictionaries, WordNet is organized by semantic relations between synsets. In this work, we use the Arabic WordNet AWN (*Arabic WordNet—Global WordNet Association*, 2013) alongside the NLTK module to extract our semantic features based on some relations such as synonyms and antonyms between essay sentences. First, we counted number of matched words not only between two adjacent sentences but also all essay sentences even matched words within the same sentence (F162) to evaluate the coherence of the essay. However, as a sentence can be expressed using different synonyms, we counted the number of synonyms in the entire essay (F203) and between all sentences using Arabic WordNet (AWN). These features also were applied over essay paragraphs to measure the similarities between essay parts (F163–F168). Moreover, we used one of AraVec models (Bakr, Mohammad, Eissa, & El-Beltagy, 2017), that proposed Arabic Word Embedding for use in Arabic NLP. For each sentence, we counted the similarity between its words and all other sentences words (F169).

E. Discourse features

As some criteria require examination between essay parts, such as coherence criterion, we propose the following features:

Arabic discourse connectives (F170–F188): we counted the number of connectives (F170) according to the list of unambiguous discourse connectives (Alsaif, 2012) in terms of discourse function, so that at least 90% of their occurrences in the Leeds Arabic Discourse Treebank (LADTB), were annotated as discourse connectives. Furthermore, we counted number of unique connectives in the overall essay (F171). We also distributed these two features over essay paragraphs and sentences (F172–F179). In addition, for each paragraph, we counted the ratio of connectives (F180) and unique connectives to the paragraph's words (F181). We then counted the ratio of number of words located between two discourse connectives to the number of connectives per paragraph in the essay (F182). In the same way, we counted the ratio of connectives (F183) and unique connectives to the sentences' words (F184). Moreover, we counted the number of times punctuation was not followed by a connective or conjunction (F202). Using the POS tag provided by MADAMIRA, we checked the number of words with conjunction tags in the overall essay (F185) as well as the number of unique conjunctions (F186). Likewise, we applied the same connectives features but with conjunctions tools in (F187) and (F188).

5 Experiments and Results

All experiments in this study were carried out using the WEKA tool (Witten et al., 1999), based on a tenfold cross-validation for the entire dataset. We built a specific model for each criterion: spelling, coherence, structure, punctuation marks, and style. Then, we computed the models' results to predict the overall score. During each model development, we considered the size of our dataset, the appropriate features, and the number of features to achieve the most accurate model with the minimum number of features possible to avoid overfitting problems (Ying, 2019). It is worth noting that we only included the effective features in each model rather than including all features.

To evaluate the system's performance, we used accuracy "Acc," which refers to the number of essays correctly evaluated by the system, as well as

Pearson’s correlation *r* to measure the relationship and association between manual scores and system scores (Benesty, Chen, Huang, & Cohen, 2009). As our dataset contains fractions, and to align with scoring in similar studies (Alghamdi et al., 2014; Azmi, Al-Jouie, & Hussain, 2019; Alqahtani & Alsaif, 2019), we considered a threshold value ***t*** in our results. Therefore, an essay is considered as correctly evaluated if the difference between the manual score and the system score does not exceed ***t***. Alghamdi et al., (2014) set ***t*** to be approximately 17% of the overall score, whereas Azmi, Al-Jouie, and Hussain (2019) set ***t*** to be 25% of the essay score. In our case, although our dataset contains many fractional numbers, we will show the results when ***t*** = 17% and ***t*** = 25% in specific criteria scores and the overall score. Table 1 shows score distributions in our dataset and the threshold values.

Criteria	**score**	**Threshold at *t*=17%**	**Threshold at *t*=25%**
spelling	4	0.68	1
structure	4	0.68	1
coherence	4	0.68	1
Punctuation marks	2	0.34	0.5
style	2	0.34	0.5
Overall score	16	2.72	4

Table 1. Score distributions and the threshold value per criterion.

A. Spelling model

We attempted many different features in the spelling model. However, the effective features were found at the surface level, lexical level, syntactical, and morphological levels as follows: (F1–F23), F29, F100, F101, F114, F117, F123, F133, F200, and F201, while the significant features were based on features related to the FARASA spellchecker (F102–F105). Using this model, the number of essays that were evaluated correctly in spelling represent 58% of our dataset when ***t*** = 17%, while it increased to 77% when ***t*** = 25%, as shown in Table 2. The variant between these two results returns to the value of the threshold and the number of fractions scores in spelling in our dataset. Further, the model achieved 0.65 when ***t*** = 17% and 0.72 when ***t*** = 25% in correlation *r* to manual evaluation. However, since the model is almost based on FARASA features, we analyzed this tool over our dataset by detecting how many times FARASA detects actual mistakes and corrects them in the right way; how many times FARASA detects actual mistakes but does not correct them as what should be; and how many times FARASA detects a correct word as a mistake, which were 2130, 85, and 250 cases, respectively. However, there were 120 words corrected by FARASA only because missing spaces in some cases that usually difficult to detect by humans such in *mAh*A/what* / ماهذا.

B. Structure model

Since a well-structured essay should contain four parts; title, introduction, body and conclusion, we included surface features (F3) and (F16) that refer to the number of paragraphs and check if the first paragraph is less than or equal to 10 words. Also, we include lexical features (F124 and F126) which related to check the existence of some keywords usually used in the introduction and conclusion parts. This model achieved 78% in Acc and 0.74 in correlation *r* when ***t***=17% and 91% in *Acc* and 0.86 in correlation *r* when ***t***=25 %. The significant feature was (F16) which refers to check if the first paragraph less than or equal 10 words which assists to indicate the title of the essay.

C. Coherence model

Coherence criterion used to evaluate the extent to which essay parts is related to the title, cohesion between essay parts, using the appropriate discourse connectives and diversity in connectives. Therefore, we included surface features (F3, F19, F20, F6–F10, F15, F12, F13), lexical features (F124 and F125) and syntactic features (F23–F99). Most of these features are generic and hold information about essay parts (lengths, general syntactic characteristics). These features are utilized to figure out essay structure which in turn assists to predict the extent of the appropriate coherence. Furthermore, we include discourse features (F170, F171, F202, F179, F137, F177,

Criteria	Level of features included per criterion	***t*** **= 17%**		***t*** **= 25%**	
		Acc	*r*	*Acc*	*r*
Spelling	Surface+Lex+(Syn and Morph+Spelling)	58%	0.65	77%	0.72
Structure	Surface +Lex	78%	0.74	91%	0.86
Coherence	Surface+Lex+Syn and Morph+Sem+Disc	79%	0.65	87%	0.69
Punctuation marks	Surface+(Lex/punctuations)+Syn/pos+Disc + Sem	75%	0.53	93%	0.74
Style	Surface+Lex+Syn+Disc+Sem	65%	0.57	78%	0.65
Overall score	A combination of the essay evaluation results in the previous criteria	90%	0.82	96%	0.87

Table 2. Features levels used for criteria modeling with their results and the overall score considering the threshold.

F187, F180–F182) since they related to connectives between essays parts in addition to the semantic features presented in (F62, F163, F169, F168 and F203) to prevent relying on only the matched words.

This model achieved 79% in accuracy and 0.65 in *r* correlation in case of ***t*** = 17% while it increased to 87% and 0.69 in accuracy and correlation *r* respectively when ***t*** = 25%. However, in some cases detecting cohesion automatically is very difficult especially if the unrelated idea is expressed within a short sentence.

D. Style model

As essay with a good style does not include repeated words without using synonyms and does not contain informal words while there is a good choice of words and diversity in the length of sentences (Ibrahim, 2006). We included surface features (F1, F5, F4, F21, F10), which predominantly investigate the length of paragraphs and sentences, lexical features from (F123, F128–F133, F134–F142) that check punctuation use, and the number of words that may affect the style. We also check morphological features (F117) as they also affect the form of words, which in turn sometimes leads to unknown or informal words. Additionally, we include discourse features and connectives (F170–F180) since punctuation might be omitted by a writer, hence there are no indications of paragraph and sentence lengths. Also, we add semantic features (F162–F168 and F203) to investigate synonymous words. However, discourse and semantics have the most impact in score prediction where the significant features were the number of discourse connectives (F170) and the number of synonyms in the whole essay (F203). The style achieved 65% in *Acc* and 0.57in correlation when ***t***=17% while it increased to 78% in accuracy and 0.65 in correlation when ***t***=25% as shown in Table2.

E. Punctuation marks model

We included surface features (F6–F9 and F12), general syntactic features such as POS (F23–F29, F31–F58) and features related to punctuation marks which refer to correct, wrong, and missing use for all marks (F134–F161). Also, we included discourse connectives (F202, F184, F187), as they separate the essay to sentences/clauses and that may assist to detect some types of mistakes such as the omission of placing comma. As each punctuation mark has certain purposes (e.g., a period used at the end of a sentence, or a comma used within a sentence to separate it into clauses), we tried to figure out the semantic impact by including features (F162–F168) which refer to the similarity between sentences. We noted that the most two significant features were (F202) that refers to the number of punctuations not followed by conjunction or discourse connectives and (F184), which refers to the ratio of unique connectives to paragraph length. The punctuation model achieved different results due to the threshold value, as in Table 2.

In the overall score, we combined the results obtained by all the previous criteria as predicted by the models, then we calculated accuracy which was 90% and 96% when ***t*** = 17% and ***t*** = 25% respectively.

In comparison to the similar studies (Alghamdi et al., 2014) and (Azmi, Al-Jouie, & Hussain, 2019), our system utilizes a wide range of features to evaluate the common criteria and can provide an evaluation of a specific criterion further to the overall score. It also does not need to train on representative-domain essays. Unfortunately, we cannot provide an accurate comparison because their datasets cannot be accessed. In addition, our system applied to a larger dataset than that used in Alqahtani and Alsaif (2019) and it supports semantic aspects which not covered by their system.

6 Conclusion

This paper introduced an Arabic essay evaluation system based on the SVR algorithm and features from different linguistic levels. Separately, we conducted experiments to predict five criteria scores; spelling, structure, coherence, style, and punctuation marks. The essay holistic score was assigned by a combination of the previous criteria scores. The experiments conducted on our dataset consisted of 200 essays. In the overall evaluation, the proposed system achieved 96% accuracy and 0.87 in correlation with manual evaluation, while it achieved 77%, 91%, 87%, 78%, and 93% in accuracy for spelling, structure, coherence, style, and punctuation marks, respectively. In the future, we look forward to expanding our dataset as our system performance improved by increasing the dataset size from 100 essays to 200 essays. Furthermore, we intend to involve some criteria that have been annotated by humans but not yet automated, such as grammar. Similar studies in foreign languages have had promising results by applying deep learning algorithms while it is unexplored in AES for Arabic writing. Therefore, we believe it is worth to apply deep learning algorithms utilizing the features extracted in this work.

References

Al-Shalabi, E. F. (2016). An automated system for essay scoring of online exams in Arabic based on stemming techniques and Levenshtein edit operations. *International Journal of Computer Science Issues*, *13*(5), 45–50.

Alghamdi, M., Alkanhal, M., Al-Badrashiny, M., Al-Qabbany, A., Areshey, A., & Alharbi, A. (2014). A hybrid automatic scoring system for Arabic essays. *AI Communications*, *27*(2), 103–111. https://doi.org/10.3233/AIC-130586

Ali, K. (2018). Inna and Its Sisters. Retrieved from https://arabicblog.info/inna-and-its-sisters

Ali, K. (2018). Kana and Its Sisters. Retrieved from https://arabicblog.info/kana-and-its-sisters

Ali, K. (2019). Fi'l Mudhari. Retrieved from https://arabicblog.info/fil-mudhari

Alikaniotis, D., Yannakoudakis, H., & Rei, M. (2016). Automatic Text Scoring Using Neural Networks. https://doi.org/10.18653/v1/P16-1068

Alqahtani, A., & Alsaif, A. (2019). Automatic Evaluation for Arabic Essays: A Rule-Based System. In *2019 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)* (pp. 1–7). https://doi.org/ 10.1109/ISSPIT47144.2019.9001802

Arabic WordNet—Global WordNet Association. (2013). Retrieved from http://globalwordnet.org/resources/arabic-wordnet/

Alsaif, A. (2012). Human and Automatic Annotation of Discourse Relations for Arabic (Doctoral dissertation, University of Leeds, Leeds, England). Retrieved from http://etheses.whiterose.ac.uk/3129/

Al-Saif, A., Alyahya, T., Alotaibi, M., Almuzaini, H., & Algahtani, A. (2018). Annotating Attribution Relations in Arabic. *LREC*.

Attia, M., Pecina, P., & ASamih, Y. (2012). Improved Spelling Error Detection and Correction for Arabic, *4*(December), 103–112.

Attia, M., Pecina, P., Toral, A., Tounsi, L., & Genabith, J. Van. (2011). A Lexical Database for Modern Standard Arabic Interoperable with a Finite State Morphological Transducer, 98–118.

Awad, M., & Khanna, R. (2015). Support Vector Regression. In *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers* (pp. 67–80). Berkeley, CA: Apress. https://doi.org/10.1007/978-1-4302-5990-9_4

Azmi, A. M., Al-Jouie, M. F., & Hussain, M. (2019). AAEE - Automated evaluation of students' essays in Arabic language (July). https://doi.org/10.1016/j.ipm.2019.05.008

Bakr, A., Mohammad, S., Eissa, K., & El-Beltagy, S. R. (2017). AraVec : A set of Arabic Word Embedding Models for use in Arabic NLP ScienceDirect (November). https://doi.org/10.1016/j.procs.2017.10.117

Benesty, J., Chen, J., Huang, Y., & Cohen, I. (2009). Pearson Correlation Coefficient. In *Noise Reduction in Speech Processing* (pp. 1–4). Springer: Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-00296-0_5

Buckwalter, T. (2004). Buckwalter Arabic morphological analyzer version 2.0. Linguistic Data

Consortium, University of Pennsylvania, 2002. LDC catalog no.: LDC2004l02. Technical report, ISBN 1-58563-324-0.

Burstein, J. (2003). The e-rater scoring engine: Automated essay scoring with natural language processing. *Automated Essay Scoring: A Cross-Disciplinary Perspective*, 113–121.

Chung, G., & O'Neil, G. (1997). Methodological Approaches to Online Scoring of Essays, CSE Technical Report 461, University of Southern California CRESST, December, *Center for the Study of Evaluation, CRESST*, *1522*(310).

Dikli, S. (2006). An Overview of Automated Scoring of Essays. *Journal Of Technology Learning And Assessment*, *5*(1), 2006–12. Retrieved from http://www.jtla.org

Dong, F., & Zhang, Y. (2016). Automatic Features for Essay Scoring—An Empirical Study. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, *1966*, 1072–1077. https://doi.org/10.18653/v1/D16-1115

Elliot, S. (2003). IntelliMetric: From here to validity. in M. Shermis and J. Burstein (eds.) *Automated Essay Scoring: A Cross-Disciplinary Perspective*, Routledge, 71–86. DOI: https://doi.org/10.4324/9781410606860

FARASA: Advanced Tools for Arabic. (2019). Retrieved from http://qatsdemo.cloudapp.net/farasa/

Gomaa, W. H., & Fahmy, A. A. (2014). Automatic scoring for answers to Arabic test questions. *Computer Speech and Language*, *28*(4), 833–857. https://doi.org/10.1016/j.csl.2013.10.005

Ibrahim, M. (2006). Looks at the technical article in modern Arabic literature (نظرات في المقال الفني في الأدب العربي الحديث). pp.21–22.

Janda, H. K., Pawar, A., Du, S., & Mago, V. (2019). Syntactic, Semantic and Sentiment Analysis: The Joint Effect on Automated Essay Evaluation. *IEEE Access*, *7*, 108486–108503.

Kukich, K. (2000). Beyond automated essay Scoring. IEEE Intelligent Systems and Their Applications. https://doi.org/10.1109/5254.889104

Lemaire, B., & Dessus, P. (2001). A System to Assess the Semantic Content of Student Essays. *Journal of Educational Computing Research*, *24*(3), 305–320. https://doi.org/10.2190/G649-0R9C-C021-P6X3

Nahar, I. K. M., & Alsmadi, I. M. (2009). The Automatic Grading for Online exams in Arabic with Essay Questions Using Statistical and computational linguistics Techniques, *1*(2), 215–220.

Pasha, A., Al-Badrashiny, M., Diab, M., Kholy, A. El, Eskander, R., Habash, N., … Roth, R. M. (2014). MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. *Proceedings of the 9th Language Resources and Evaluation Conference (LREC'14)*, 1094–1101.

SAFAR: Software Architecture For ARabic. (2013). Retrieved from http://arabic.emi.ac.ma/safar/?q=examples#

Shehab, A., Faroun, M., & Rashad, M. (2018). An Automatic Arabic Essay Grading System based on Text Similarity Algorithms, (April). https://doi.org/10.14569/IJACSA.2018.090337.

Surya, D., Madala, V., Krishna, S., Surya, D., Madala, V., Gangal, A., … Sureka, A. (2018). An empirical analysis of machine learning models for automated essay grading.

Witten, I. H., E. Frank, L. E. Trigg, M. A. Hall, G. Holmes and S. J. Cunningham. (1999). Weka: Practical machine learning tools and techniques with Java implementations. Proc ICONIP/ANZIIS/ANNES99 Future Directions for Intelligent Systems and Information Sciences, pp. 192–196.

Ying, X. (2019). An Overview of Overfitting and its Solutions: IOP Conf. Series: Journal of Physics: Conf. Series 1168. 2019.