Using Automated Procedures to Score Written Essays in Persian: An Application of the Multilingual BERT System

by

Tahereh Firoozi

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Measurement, Evaluation, and Data Science

Department of Educational Psychology

University of Alberta

Abstract

The automated scoring of student essays is now recognized as a significant development in both the research and practice of educational testing. The majority of the published studies on automated essay scoring (AES) focus on outcomes in English. Studies on multilingual AES—meaning languages other than English—are, by comparison, practically non-existent. The purpose of this study is to develop, describe, and evaluate the first AES system for scoring essays in the Persian language using multilingual BERT. Multilingual BERT is a transformer-based encoder model for language representation that uses an attention mechanism to learn the contextual relations between words and sentences in a text. The Persian language version of BERT was used to grade 2,000 holistically-scored essays written by non-native language learners in Iran on a scale that ranged from 1 (Elementary) to 5 (Advanced). The performance of the BERT AES model was examined against a baseline model that only included a word embedding layer (Word2Vec). The models were evaluated using four metrics: the Quadratic Weighted Kappa, the Kappa coefficient, model accuracy, and error analysis. The BERT AES model performed with high classification consistency (QWK=0.84 vs. Baseline QWK=0.75; κ= 0.93 vs. Baseline κ= 0.82). The result from the accuracy measures shows that the BERT AES model correctly scored about 73% of the total number of essays. Of those essays considered correctly classified by the BERT AES system, more than 70% in each level except for Advanced were scored the same by the human raters (i.e., true positive). Among the essays that were incorrectly classified, more than 70% in each score level—except for Advanced—were considered incorrect (i.e., true negative). Error analysis showed that each level had some overlap with the adjacent levels, with the Upper-Intermediate and the Advanced levels having the highest number of overlaps. These results demonstrate that the BERT AES model can be used with a

high degree of accuracy to predict the essay scores produced by the raters in this study. The one area where the performance results were comparatively weak was at the Advanced level due to the smaller number of essays (*n*=238). Augmentation provides a method that can be used to solve the text data sparsity problem when using low-resource languages like Persian. To improve model performance, sentence-level data augmentation was implemented by adding 20% more data to each score level. This approach improved the classification performance of the BERT AES model ($QWK_{Pre\text{-}SLDA} = 0.84$ vs. $QWK_{post\text{-}SLDA} = 0.96$; $\kappa_{Pre\text{-}SLDA} = 0.88$ vs. $\kappa_{post\text{-}SLDA} = 0.96$) thereby demonstrating the benefits of text augmentation. The architecture and methods described in this study can be easily adapted and used to score essays written in other non-English languages, thereby supporting the application and wide-spread use of multilingual AES.

Acknowledgements

I would like to express my deepest appreciation and gratitude to my supervisor Dr. Mark J. Gierl, for his guidance, support, and encouragement throughout my doctoral journey. His knowledge, expertise, insights, and patience have been invaluable in shaping my research and helping me develop as a scholar and as a human. Dr. Gierl has not only been my academic supervisor, but he has also helped me to know myself better and peruse my goals in life. Dr. Gierl, your mentorship is a light to my future life journey.

I am also grateful to the members of my dissertation committee Dr. Okan Bulut and Dr. Hollis Lai for their insightful feedback and constructive criticism, which helped me refine my ideas and approach.

I would also like to thank my external examiners Dr. Ying Cui at the University of Alberta and Dr. Gregory J. Cizek at the University of North Carolina for providing constructive criticism and positive feedback that affected the improvement of the current version of this study.

I would also like to express my acknowledgment to the Saddi Foundation that provided me with the Persian Essays dataset for this study.

I would like to thank my colleagues and friends in the Centre for Research in Applied Measurement and Evaluation (CRAME) who provided me with a supportive and stimulating academic environment, and for their encouragement and advice during my PhD program.

Finally, I would like to express my heartfelt appreciation to my parents, Rakhshaan and Ali, my sisters, Fatema and Zahra, and my lovely niece, Sara, for their unwavering love, support, and encouragement. Their sacrifices and belief in me have been the driving force behind my academic pursuits.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

INTRODUCTION

Extraordinary learning opportunities are now available through instructional technologies that permit students to access massive open online courses and other online learning programs. For example, the World Economic Forum claimed that 21 million students registered for Coursera's online courses in 2016. The pandemic only served as a catalyst for the migration to online learning as the number of registrations skyrocketed. Coursera enrollment increased more than three-fold to 71 million in 2020, with an additional 21 million registrations in 2021, bringing the most recent count to 92 million (World Economic Forum, 2022). This example demonstrated that students have abundant opportunities to access online learning resources and are capitalizing on these opportunities.

One significant challenge to be addressed in online teaching and learning is the development and administration of educational tests. In particular, administering written-response assessments that yield valid and reliable test score interpretations poses an important challenge because of the *scoring process*. Written-response assessments such as essays are typically evaluated by a human rater to yield inferences about the student's knowledge, skills, and competencies. The traditional method for scoring involves training a human rater to interpret and apply a rubric that can be used to score the students' responses. Unfortunately, human scoring is time-consuming because it requires human raters to evaluate a large number of essays. It is also expensive because it requires extensive logistical efforts to hire human raters, train them to consistently interpret and apply the scoring rubric, and deploy them to evaluate each student's written-response task. However, an even more significant challenge remains to be solved. Using human raters to evaluate written-response assessments is virtually impossible to scale in a timely and cost-

effective manner. If, for instance, 100 students complete a written-response assessment such as an essay, then 100 essays must be scored by the human raters. This scale is reasonable with enough trained raters. If 92 million Coursera students complete an essay as part of their online courses, then 92 million essays must be evaluated by raters. This scale is unreasonable because the time and expense required to train a legion of human raters who must then score the essays would hinder a real-life application of human scoring.

One way to address this scaling challenge is to implement automated essay scoring (AES) so that machines can be used to help humans score students' written-response assessments. AES can be described as using computer algorithms to score unconstrained open-ended written tasks by having a computer mimic the human raters (Shermis, 2014). Research on the development and application of AES has become increasingly important in the last decade as practitioners attempt to implement methods that can be used to efficiently and accurately score students' written-responses at scale. The need for these methods has only been amplified in the past three years as students migrate to online learning environments en masse, resulting in new scoring practices that allow educators to evaluate large numbers of written-response assessments efficiently and economically.

Overview of Automated Essay Scoring

AES is the process of evaluating and scoring written text with computer programs. An AES system is a computer program designed to evaluate student responses so that the program yields scores that are similar to those of trained human raters (Shermis, 2014). AES is a statistical classification method where input linguistic features in the text are mapped to a specific output, like an essay score, so that the input and output are related to one another statistically. The mapping function is called a scoring model, and it can be used to classify new instances of the

input text into the output score. The use of a scoring model allows educators to scale the assessment because instead of a human, the computer can be used to score students' written tasks. To emulate human scoring, the AES program builds the model using techniques and procedures from the fields of natural language processing and computational linguistics where features are extracted from the example instances, called the training dataset, that have been scored by human raters. The AES method is described as supervised when a training dataset is used. A supervised machine learning algorithm uses training samples drawn from a written-response prompt that contain scores from raters. The algorithm analyzes the patterns in the scored essays in order to model the behaviour of the human raters. This step yields statistical weights that are used by the scoring model in order to classify a new set of written-response tasks produced by a different group of students.

Automated Essay Scoring: A Four-Step Process

The AES process can be described in four steps (see Gierl et al., 2014). The first step is pre-processing. Pre-processing requires the student response data to be available in a format that can be processed using an AES system. In an online learning system, students' data are available immediately in an electronic form. Transformations and modifications are conducted on the pre-processing data so that raw data from the students and raters is annotated into a data format that an AES system can read. The outcome from the pre-processing step is that raw input data from the students (i.e., essays) and the ratings (i.e., essay scores) are readable by the AES system.

The second step is feature extraction. Feature extraction requires a process where the input text is transcribed into features representing the text and can be used by the machine learning algorithm. Traditional AES approaches focus on extracting linguistic features as variables from the text that help predict the final essay score (e.g., Attali, 2013; McNamara et al., 2015;

McNamara et al., 2014; Page, 1994). The benefit of this approach is that the features are identified before the AES analysis is conducted, thereby serving as indicators of written-response quality that can be interpreted by humans. The drawback of this approach is that the model may not produce a high level of accuracy in predicting human scores (e.g., Shin & Gierl, 2020).

To address the limitations of the traditional scoring approaches, alternative AES approaches can be used to maximize the predictive accuracy of reproducing the final essay score. These modern AES approaches automatically extract features to model the association between each written essay and the final essay score, where the goal of feature extraction is to maximize the predictive accuracy of the scoring model (Dong et al., 2017; Kim, 2014; Mikolov et al., 2010). For instance, deep learning is a modern feature extraction method designed to maximize the predictive accuracy of the scoring model.

A traditional AES approach begins by manually extracting relevant features from the text. These features produce a model that categorizes important text features. A modern AES approach using deep learning begins by automatically extracting relevant text features. The extraction process requires end-to-end learning, which means that a neural network is given a text and the data with the rater's scores. The task for the network is to learn how to reproduce the scores as output using the text as input. The word "deep" is a reference to the hidden layers in the network. Traditional networks often have 2-3 hidden layers. By way of comparison, deep networks may have 100. This large number of layers helps the network learn minute details that, in turn, are used to maximize the score prediction. The strength of the modern AES approach is that the model can predict the essay scores with high accuracy. The weakness of a modern AES approach is that the complex feature structures used to maximize the predictive accuracy of the scoring model are difficult to interpret. In other words, the modern approach can accurately

predict scores from humans using variables that are difficult to interpret linguistically. Modern approaches also require larger samples of essays compared to traditional approaches.

The third step is creating a scoring model using machine learning algorithms. A scoring model contains a list of the extracted features from the second step. This model contains the input mapped onto the human rater scores that serve as the output so that a clear relationship exists between the input and output. Machine learning algorithms learn the relationship between the text feature input and essay score output by evaluating responses in the training sample. Different algorithms can be implemented to learn the classification process. Machine learning is an analytic process where the feature input is mapped onto the score output with the purpose of creating a text classifier that can be used to score the written responses of students. The algorithm must learn how to describe the scoring function by analyzing student and rater data in the training set.

The fourth step is essay scoring. The machine learning model is used to score written responses using the same essay prompt but with a different group of students. This group is called the validation sample. An AES system that contains a model that scores data from one existing essay is called prompt-specific. An AES system that contains a model that scores data from two or more essays intended to be interchangeable with one another is called generic. Most AES studies used in operational testing programs use prompt-specific scoring models because they yield score predictions that tend to be more accurate than the generic models.

After the students' written responses are scored, different performance measures are used to evaluate the scoring model. Model validation is the process of comparing the predictions produced by the AES system and the human raters (Attali, 2013; Chung & Baker, 2003; Williamson et al., 2012). Human raters are the "gold standard" in this comparison, thereby

serving as the criterion for measuring the accuracy of the AES system. Two commonly used measures of score agreement are Kappa and Quadratic-Weighted Kappa (QWK). Kappa is the agreement between the raters and the AES scores that includes a measure of chance. Kappa provides a chance-corrected statistic that represents the proportion of observed agreement to the maximum agreement corrected for chance agreement (Siegel & Castellan, 1988). Kappa, however, does not account for the degree of disagreement. Therefore, a weighted Kappa score called QWK addresses this limitation. A QWK of 0.80 or higher indicates strong agreement (Williamson et al., 2012).

Benefits of Implementing Automated Essay Scoring

AES yields pedagogical and economic benefits when scoring students' written-response assessments in online learning environments (e.g., Klebanov & Madnani, 2022; Kumar & Boulanger, 2020). AES has many applications that range from scoring formative written-response tasks to summative high-stakes essays. The pedagogical benefits of AES and the optimism for automated scoring have a long history in educational testing. Ellis Page, in 1966, argued that daily writing tasks are critical for student learning, mainly when the student is provided with immediate feedback to monitor the learning progress. He claimed:

> *Almost everyone knows what students should do in English: preferably a daily writing stint; at least a weekly theme. Each exercise should be carefully planned, and each should be returned to the student with extensive and wise comment, correction, exhortation, and encouragement, thoughtfully managed by a teacher who understands the student's fumbling progress from inarticulate solecism toward organized clarity.* (Page, 1966, p. 238)

Given that providing students with immediate feedback on their written tasks is extremely time consuming and laborious for teachers, the use of computer for scoring was considered essential to make the ideal learning situation feasible:

> *Just for a moment, then, imagine what the result would be if all student essays could be turned over to a computer, which would perform a stylistic and subject-matter analysis according to the general rules desired, and deliver extensive comment and suggestion for the student to the teacher by the first bell the next day.* (Page, 1966, p.239)

With the use of deep learning models in AES, Page's dream of automated scoring came true.

In addition, artificial intelligent (AI) systems are now available that can provide feedback to students within seconds of receiving their responses (e.g., Kumar & Boulanger, 2020). These feedback systems can be quite transparent through explained AI models so that the teacher can assess the suitability of an agent for the job and hire the candidate agent that comes closest to describing human performance. Explanations derived from these models could therefor serve as formative feedback to the students (Arrieta et al., 2020).

Another pedagogical benefit of AES is that it makes the scoring process more reliable. AES eliminates the source of human subjectivity in scoring that is usually caused by rater fatigue, rater's expertise, severity, leniency, scale shrinkage, stereotyping, Halo effect, rater drift, perception difference, and inconsistency (Klebanov & Madnani, 2022).

Finally, AES is a scalable method thereby allowing educators the important benefit of evaluating large numbers of written-response assessments efficiently and economically. Assessing students' written tasks in large summative tests necessitates the recruitment of a vast number of human raters that must be trained to implement a scoring rubric reliably. Particularly in high-stakes written tasks where the prompts should be updated from test to test, implementing

and updating the scoring rubric and consequently re-training a high number of human raters is extremely challenging. AES provides the benefit of a more cost-efficient scoring process with comparable reliability to that of human raters (Zhang, Williamson, Breyer, & Trapani, 2012).

Background of the Problem

To date, the majority of the published AES studies have focused on essays written in English (Ramesh & Sanampudi, 2021). Studies on multilingual AES—meaning essays written in languages other than English—are, by comparison, practically non-existent except in a small number of languages such as Japanese (Hirao et al., 2020), Chinese (Li & Dai, 2020), and Arabic (Shehab et al., 2016). Multilingual AES is a critically important research area because the language of assessment for many students throughout the world is not English. For instance, the most popular achievement test in the world—the Programme for International Student Assessment (PISA)—is administered in over 90 different languages (OECD, 2021) to encompass the breadth of languages used to instruct students around the world.

Persian is an Indo-European language spoken by more than 110 million people and is an official language in Iran, Tajikistan, and Afghanistan. In terms of script, Persian is similar to Semitic languages (e.g., Arabic). Linguistically, Persian is an Indo-European language (Masica, 1993) and thus distantly related to most of the languages of Europe and the northern part of the Indian subcontinent. These attributes make Persian a unique case to study in terms of language technologies. Although Persian is a widely spoken language (Simons & Fennig, 2017), automated text analysis on this language remains limited because Persian is a low-resource language in terms of data and computational linguistic tools (e.g., Habib, 2021; Khashabi et al., 2021). In other words, Persian datasets are not readily available, and little research using computational linguistic methods in Persian has been documented in the literature.

The unique writing and linguistic system of the Persian language also necessitates the development of a specific and unique AES system. The alphabetic system of Persian is different from English, meaning that the majority of the NLP libraries currently in use are not applicable to Persian. Furthermore, Persian's standard orthography uses a combination of spaces and semi-spaces (zero-width non-joiners), which are often ignored or confused, leading to orthographic inconstancies and added sparsity. In standard Persian orthography, semi-space characters show inter-word boundaries. Around 8% of all tokens in the Persian dependency treebank have semi-spaces (Rasooli et al., 2013). Another noticeable difference in Persian orthography is that, unlike English which is left to right, Persian is right to left.

Persian has a complex morphological system that contains a heavily suffixing affixational morphology with no expression of grammatical gender (Amtrup et al., 2000). Suffixes in Persian are a source of challenge in automated text analysis. For example, the suffix 'ک' (pronounced /K/) can have three different meanings when added to the end of a noun such as, 'دختر' (means girl) → 'دخترک' (pronounced /Dokhtark/) means 'a lovely girl', 'an inferior girl', and 'a small girl' depending on the context of use. Another example is the suffix /ی/ (pronounced /i:/) that as a versatile suffix can convert nouns to adjectives (e.g., ایران [Iran] →ایرانی [Iranian]), adjectives to nouns (e.g., سبز [green]→سبزی [vegetable]), and infinitives to adjectives (e.g., دیدن [to see] → دیدنی[worth seeing]). Adjectives in Persian have a limited inflection space: they may be simple, comparative, or superlative. In comparative and superlative forms (except for Arabic loan words), a suffix attaches to the adjective: '+تر (pronounced/tar/=er [in English]) for comparative and '+ترین (pronounced/tari:n/=est [in English]) for superlative adjectives. English uses both suffixes ('+er/+est') and multi-word construction with 'more/most', in addition to some irregular

cases such as ‘good’, ‘better’, and ‘best’. As such, it might be hard to define a consistent pre-process scheme for adjectives in Persian with respect to English.

Verbs in Persian may be inflected in different combinations for tense, mood, aspect, voice, and person. For example, the past tense stem is used with another auxiliary verb to create the future form. When an auxiliary verb is used, prefixes attach to the auxiliary verb instead of the root. The negative marker ‘+ن ‘ (pronounced /n/ [means ‘not’]) and the object pronouns are attached to the verbs, leading to more than 100 verb conjugated forms (Rasooli et al., 2013). For example, the verb نمیخواندمش’ (pronounced /n.my.xwAnd.m.sˇ/) can be tokenized to /n/+ /my/+ /xwAnd/ +/m/ +/sˇ/ means ‘I was not reading it’ [lit. ‘not+ was(continuous)+ read(past) +I +it’]. This example shows that Persian is a pro-drop language where almost half of the verbs in the Persian dependency treebank do not have an explicit subject (Rasooli et al., 2013).

To address the challenges inherent to scoring a Persian written-response task, a large language-specific dataset is needed to train an AES system in order to learn the varieties in the essays. In AES, training data are the essays that are graded by human raters. These graded essays (also called labeled data) help machine learning algorithms to learn how the essays should be scored. When there is a complex morphological system, which is characteristic of Persian, the AES system needs to see a broad range of content in the training data in order to learn how to score the essays that are not evaluated by human raters (also called unlabeled data). Unfortunately, Persian is also one of the low-resource languages meaning that few free and readily available language resources with labeled data are available for use in AES applications.

The limited number of studies on automated text analysis that have been conducted in Persian showed that the text classification task is very challenging because of the sparsity of the available labeled texts. Research has also demonstrated that the use of deep learning models can

address some of the existing problems in low-resource languages, such as Persian (e.g., Roshanfekr et al., 2017). For instance, Roshanfekr et al. conducted a study on sentiment analysis using electronic products based on customers' review feedback in Persian. They argued that language representation is one of the main challenges in automated text analysis of the Persian language. To address this problem, they implemented vector space pre-trained models over vocabulary indexing. They argued that the use of vocabulary indexing is challenging for text analysis in Persian because there is a vast number of derivational affixes in the language. For example, the word 'می‌روم' (pronounced /Mi:.ræ.væ.m/) is given one index using traditional word representation, while it means 'I am going' where the subject 'I' is represented with suffix 'م', the tense (i.e., present progressive) 'am+ing' is represented with the prefix 'می' and the infinitive (i.e., go) is 'رو'. The results of their study indicated that using pre-trained deep learning models for text representation can solve some but not all of the indexing problems in Persian text analysis.

Despite these challenges, the problem of developing an AES system for Persian remains an important problem and should be solved because Persian is the language of instruction and assessment at schools and higher education institutions throughout the world. For example, the UNESCO Institute for Statistics reports that 64% of the population in Iran are students, and the net enrollment rate at the elementary school level is reported as high at 99.1% (UNESCO Institute for Statistics, 2015).

Purpose of the Study

The purpose of my dissertation research is to develop, describe, and evaluate the first AES system for scoring essays using the Persian language. Recent innovations in NLP methods and technologies, such as the development of transformers for language representation, provides a

novel method that can be used to solve the text analysis problem inherent to low-resource languages such as Persian. Transformers are deep learning models containing a self-attention mechanism that learns the context of input text data, thereby providing different weights to different parts of a text based on understanding the relationship between the sequence of inputs with the entire text (Vaswani et al., 2017). Using an attention mechanism for contextual understanding created the opportunity for developing pre-trained language models such as BERT (BERT stands for **B**idirectional **E**ncoder **R**epresentation for **T**ransformers), which can be used to analyze text in multiple languages. More specifically, Multilingual BERT (Devlin et al., 2018) is a transformer-based approach created by the Google AI Research team (see https://huggingface.co/bert-base-multilingual-cased). It was pre-trained in 104 languages, including Persian (Pires et al., 2019). BERT is a state-of-the-art method for many NLP tasks in English (Uto, 2021). Delvin et al. demonstrated that BERT could achieve the highest performance (i.e., accuracy score, F_1 score, and Spearman correlations) on 11 NLP tasks, including question answering and text classification. The accomplished state-of-the-art performance of BERT is supported by bidirectional training on massive amounts of data and leveraging transformers architecture (Vaswani et al., 2017) to revolutionize the field of NLP. In this study, I use multilingual BERT as a state-of-the-art method to develop the first AES system for use in Persian.

CHAPTER 2

LITERATURE REVIEW

The Mechanism of Transformer-Based Systems

The emergence of transformers (Vaswani et al., 2017) has revolutionized the field of NLP by significantly increasing the performance of several NLP tasks, such as machine translation, question answering, and language modeling (Chernyavskiy et al., 2021). Transformers are encoder-decoder-based neural networks used to solve sequence-to-sequence problems by finding a mapping function f from an input sequence (e.g., word or sentence) of n vectors $X_{1:n}$ to a sequence of m target vectors $Y_{1:m}$. That is,

$$f: X_{1:n} \rightarrow Y_{1:m}. \quad (1)$$

Before the emergence of transformers, the state-of-the-art encoder-decoder models for sequence-to-sequence tasks were deep neural network (DNN) and recurrent neural network (RNN) models. The problem with the DNN models is that they can only define a mapping where input and target sequences can be encoded with the same length (Sutskever et al., 2014), which are determined without considering of context of the texts. For example, in machine translation of English ⇋ Persian languages using DNN, the input vector 'I am going' is expected to be translated to a target vector of the same length in Persian. However, as demonstrated in Chapter 1, the morphological system of Persian is not the same as that of English, which results in the target vector of a different length—'I am going' ⇋ 'می‌روم'). This lack of flexibility can also be problematic in other text analysis tasks, such a text summarization, because the number of target vectors is conditional on the entire context of the input sequence. For example, an article of 1,000 words can be summarized into both 200 words and 100 words, depending on its content.

Recurrent neural network (RNN) models solved the problem inherent to DNN models by adding a hidden layer c to the model which auto regressively updates the conditional probability of the output vectors $Y_{1:m}$ given the hidden states of the input vectors $X_{1:n}$. In other words,

$$P_{\theta dec}(y_i|Y_{0:i-1}, c). \quad (2)$$

While RNN-based models revolutionized the Google translation engines in 2016, they have two main problems. First, RNNs suffer from the vanishing gradient problem making it very difficult to capture long-range dependencies within the text (Hochreiter et al., 2001). For instance, if a system is developed to predict the next word in a sentence, the network must have a better knowledge of the preceding words in the text for more accurate prediction. Given that in RNN, the hidden weights are updated recurrently to decrease the error function, in long texts where there are more hidden weights at different time steps, the initial weights are multiplied by the updated weight. However, because the initial weights are small, this multiplication decreases the gradient value very quickly, so the system stops training before it learns the whole text. This problem with sequential training was solved using a parallel structure in encoding the input sequence of different lengths (Vaswani et al., 2017).

To push the boundaries of RNN models, Vaswani et al. (2017) introduced transformers that avoid sequential computation and, instead, use an attention mechanism to compute sequence representation. The title of their manuscript—"Attention Is All You Need"—provided a concise and accurate summary of their method. The attention mechanism processes an input sequence of $X_{1:n}$ of variable length n in parallel that allows the model to map the input sequence to a contextualized encoding sequence, given as

$$p_{\theta dec}(y_i|Y_{0:i-1}, X_{1:n}). \quad (3)$$

The attention mechanism in transformers solved a critical problem that limited RNN models because these models were unable to capture the importance of each input in relation to other inputs across the entire text. That is, RNN models cannot capture the contextual information of the text for accurate input representation due to memory constraints when the sequence of input is long, such as in written essays. This limitation can be overcome with an attention mechanism. The attention mechanism in transformers can capture all of the contextual information within a text to calculate the weighted sum of values for each token (e.g., word) in a sequence of input (e.g., sentence). For example, in the sequence of input $X_{1:6}$=I "want to buy a car", the representation of X_4 depends not only on the input X_4= "buy", but also on all other words "I", "want", "to", "a", and "car" in the sequence. This feature allows for the modeling of global dependencies in all of the sequential inputs without regard to their distance in the input or output sequences. Hence, in encoding or decoding the representation of an input sequence, the attention mechanism allows transformers to learn the context of the input by parallelizing all the surrounding inputs within training examples (Vaswani et al., 2017).

Bidirectional Encoder Representations from Transformers

The application of transformers in NLP tasks was accelerated with the advent of a specific transformer-based language model called the **B**idirectional **E**ncoder **R**epresentation for **T**ransformers or BERT (Devlin et al., 2018). BERT is a transformer-based encoder model for language representation that uses a multi-head attention mechanism and a bidirectional approach to learn the contextual relations between words and sentences in a text for an accurate representation of the entire text. Multi-head attention is a mechanism for training transformers

that connect each input vector to all other vectors in the text in order to consider the context in word representations. The bidirectional in BERT refers to the training process where the transformers learn texts from both directions, left and right.

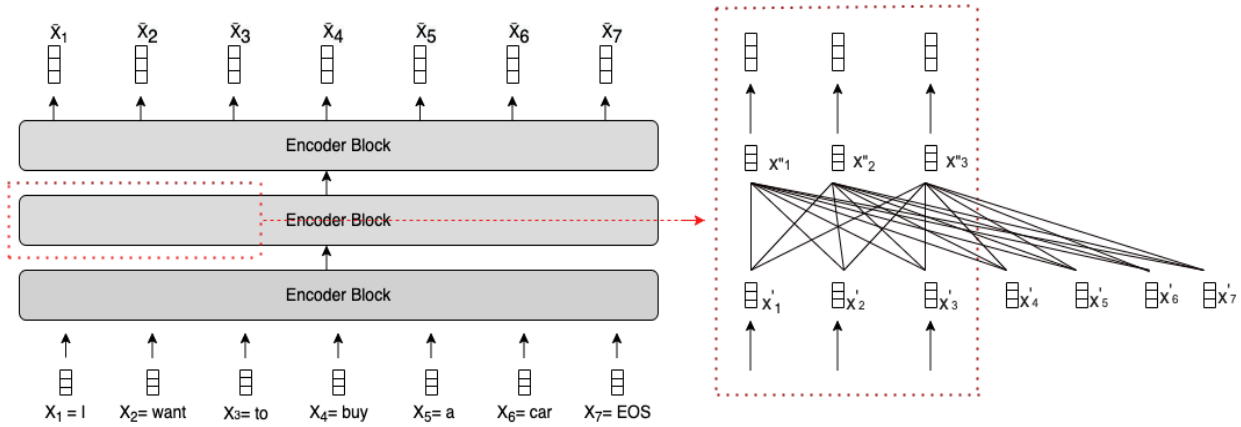
For text classification tasks like essay scoring, BERT has achieved a state-of-the-art accuracy performance level among the currently available transformer models (Uto, 2021). The exceptional performance of BERT is achieved for two reasons: 1) It uses a bidirectional approach in learning the context and it is a self-supervised learning model, and 2) unlike other language models that read the text input sequentially (i.e., left-to-right or right-to-left), the encoder in BERT reads the entire sequence of the input (i.e., words or sentences) at once. This characteristic allows the model to learn the context of the input on the surrounding input which results in a more accurate performance of the system (Devlin et al., 2018).

BERT builds on two recent trends in the field of NLP. First, the transformer model and the transfer learning. BERT contains a transformer-based encoder that is a stack of residual encoder blocks. Each encoder block consists of a bidirectional self-attention layer, followed by two feed-forward layers (Figure 1). Second, the bidirectional self-attention layer puts each input vector $X_i', \forall i \in \{1, \dots, n\}$ into a specified relation with all input vectors $X_1', \dots, X_n'$ and by doing so transforms the input vector X_j' to a contextual representation of itself, defined as $X_j^{"}$. As can be seen in Figure 1, each output vector of the self-attention layer $X_i^{"}, \forall i \in \{1, \dots, 7\}$ depends directly on all input vectors $X_i', \forall i \in \{1, \dots, 7\}$. This means that, for example, the input vector representation of the word "buy" (i.e., X_4') is put into direct relation with the word "want" (i.e., X_2'), but also with the word "I" (i.e., X_1'). The output vector representation of "buy" (i.e., X_4') therefore represents a more refined contextual representation for the word "buy". With BERT, the training procedure to obtain the global dependency for representing the input vectors is

bidirectional meaning it reads from left to right and from right to left. For instance, the model is trained to learn the sequence from the left to right "$X_1 = I,\ X_2 = want, X_3 = to,\ X_4 = buy,\ X_5 = a,\ X_6 = car, X_7 = EOS$" and also from right to left "$X_1 = car,\ X_2 = a, X_3 = buy,\ X_4 = to\ X_5 = want,\ X_6 = I, X_7 = EOS$".

Figure 1.

BERT Model Architecture



The left side of Figure 1 represents a bidirectional self-attention network. Each input vector X_i' of an input vector $X_{1:n}'$ of an encoder block is projected to a key vector k_i value vector v_i and query vector q_i through three trainable probability weight matrices W_k, W_v, W_q:

$$k_i = W_k X_i', \tag{4}$$

$$v_i = W_v X_i', \text{ and} \tag{5}$$

$$q_i = W_q X_i'. \tag{6}$$

The more similar (cosine similarity) one of the key vectors $k_1, \ldots, k_n$ is to a query vector q_j , the more important is the corresponding value vector v_j for the output vector $X_j^{"}$. So the output of the bidirectional network is encoded as:

$$X^{"}_{1:n} = V_{1:n}\, Softmax(Q^{T}_{1:n} K_{1:n}) + X^{"}_{1:n}. \quad (7)$$

The output $X^{"}_{1:n}$ is then computed via a series of matrix multiplications and a softmax operation, which can be parallelized effectively for long-range contextual representations.

The idea of transfer learning is to train a model on one task and then leverage the acquired knowledge to improve the model's performance on a related task. BERT is first trained on two unsupervised tasks. The first is masked language modeling (MLM). MLM is predicting a missing word in a sentence. In MLM, the model randomly masks 15% of the words in a sentence and then runs the entire masked sentences through the model in order to predict the masked words.

BERT also uses a second unsupervised task called next sentence prediction (NSP). NSP is predicting if one sentence naturally follows another. In NSP, the model learns to understand longer-term dependencies across sentences. Using the NSP strategy allows the model to predict each two-sentence sequence that follow one another in a text. BERT learns this knowledge by receiving masked sentence embeddings that are concatenated in pairs as inputs during pre-training. Half of the embeddings are random and the other half are actual sentence pairs from the pool of training data. For example, the model receives sentence A and sentence B to predict whether sentence B is the next sentence, or whether it is not the next sentence. This process continues and the model learns from the error rates in each prediction until it fully predicts the accurate sequence of sentences in a text.

Using the MLM and NSP algorithms has allowed BERT to solve the bottleneck problem inherent to supervised learning language models where massive amounts of labeled data are required for training (Hui et al., 2020). BERT is a self-supervised learning model that was

developed and pre-trained by Google on raw publically available data, such as millions of sentences from the internet (e.g., Wikipedia), without human intervention. Being pre-trained on a very large number of texts by Google, BERT contains 110 million parameters that are ready to be fine-tuned on a new training data. In short, BERT contains rich structural information about language that enables it to achieve a high degree of predictive accuracy on task-specific classification even with limited amounts of training data (Sun et al., 2019). This unique characteristic of BERT is beneficial for text classification tasks in different research areas, such as education, where there is simply not enough labeled data to conducted AES analyses using the conventional methods, particularly with text that is not written in English.

Recent Applications of BERT in Educational Testing

BERT has been used to solve different problems in education including the processing of log data used with online learning platforms (e.g., Chanaa & Faddouli, 2020) and with intelligent tutoring systems (e.g., Khayi, 2021). BERT has also been used to solve a limited number of problems in educational testing such as processing student's selected-response data for computerized testing (e.g., Scarlatos et al., 2022), scoring reading comprehension tests (e.g., Tseng et al., 2019), grading short-answer tests (e.g., Camus & Filighera, 2020; Haller et al., 2022; Sung et al., 2019), and grading long-answer tests such as essays (e.g., Beseiso et al., 2021; Mayfield & Black, 2020; Rodriguez et al., 2019).

One reason that BERT has been used in educational data mining, especially text analysis, is that there is often data sparsity for the variable of interest (e.g., students' score or outcome). As a result, supervised learning models cannot accurately learn the representation of the data for future predictions because there is not enough data in the training set. BERT solves this problem because it is a pre-trained model containing an abundance amount of information that can be

used to learn the representations within the data even with a small training set (Delvin et al., 2018). The performance of BERT in these educational contexts provides evidence that this method could also be used to solve some of the more challenging problems in educational testing (Condor, 2020).

To-date, BERT has been used as a contextual embedding technique for text representation (Liu et al., 2020) than can help solve AES problems such as improving score accuracy and overcoming the limitations associate with data sparsity (Wangkriangkri et al., 2020). The accuracy of AES scoring was improved with the introduction of pre-trained word embedding techniques (Firoozi et al., 2022; Klebanov & Madnani, 2022; Mikolov et al., 2013; Pennington et al., 2014). Pre-trained wording embedding is an important component of modern NLP systems where each of the words in a text is assigned a representation (i.e., feature vector) based on the ratio of co-occurrence probabilities between pairs of words in a text. The limitation of word embedding techniques that were used before the introduction of BERT is that they only obtain a single global representation for each word and do not consider the context. Therefore, all the words in the essays that have the same semantic meaning are assigned the same representation thereby ignoring the context that is unique to each essay. BERT address this problem through context embedding by encoding texts on a case-by-case basis and giving a unique representation to each text (Delvin et al., 2018). Using this approach, BERT produced a higher level of scoring accuracy compared to those AES systems where word embedding techniques, such as GloVe, were used (Rodriguez et al., 2019; Wangkriangkri et al., 2020). Moreover, the outcomes from recent studies using English texts have demonstrated that AES systems that use BERT for text representation can achieve a reliability index comparable to that of the human raters (e.g., Xue et al., 2021).

The common evaluation indices used to measure the score correlation between the human rater and the AES system are Cohen's Kappa (κ) and Quadratic Weighted Kappa and (QWK). The Kappa index measures the extent to which the scores predicted by the AES system are reliable. A Kappa of 1.0 indicate perfect agreement whereas 0.0 indicate no agreement. Kappa results over 0.80 indicate a high level of consistency for the scores (Shermis, 2014). QWK is a variant of Cohen's Kappa, with quadratic weights for misclassifications based on how close the ratings are to the correct score level using an ordinal scale. In the current study, QWK was adopted as the main evaluation metric because it can easily be used to compare the performance of our model with the performance of models used in similar studies. QWK varies from 0 (random agreement between raters) to 1 (complete agreement between raters). Typically, values between 0.60 and 0.80 QWK are used as a lower bound estimate for an acceptable reliability outcome using human raters in a high-stakes testing situation (Williamson et al., 2012).

The BERT-based AES model proposed by Xue et al. (2021) outperformed (average QWK= 0.83) the previous state-of-the-art deep learning model (Dong et al., 2017) that used GloVe word embedding technique with the Long Short-Term Memory (LSTM) model (average QWK= 0.76) on all of the essay prompts in the Automated Student Assessment Prize (ASAP) competition. The ASAP was sponsored by the Hewlett Foundation in 2012. The competition focused on the performance of AES technology for scoring student essay (Shermis, 2014). The task in the competition was for the AES systems to reproduce the essay scores created by human raters. Scores from human raters were obtained on 12,978 essays written by students at three grade levels in six American states across eight different written prompts under standardized test administration conditions. The essays were written by and equal number of male and female students from diverse ethnic backgrounds. The essays were scored by teachers who were trained

to use a rubric that was appropriate for each of the written prompts. The BERT-based AES model also achieved a higher human-computer agreement (0.78<QWK< 0.88) than the human-human agreement (0.62<QWK<0.85) (Doewes & Pechenizkiy, 2021).

Beseio et al. (2021) proposed a transformer-based AES system to improve the accuracy of the existing AES neural network models. They argued that traditional language models, such as GloVe and Word2Vec, along with RNN models that were used in the existing AES systems gave the same embedding for words used in different contexts, so they did not address text coherence in scoring. This issue can be solved using transformers coupled with deep learning models. Their proposed model containing RoBERTa was trained and tested on the ASAP dataset. RoBERTa is a version of BERT that was released by Facebook. The difference between RoBERTa and BERT is that unlike BERT that is pre-trained on a combination of MLM and NSP tasks, RoBERTa was pre-trained just on the MLM task using dynamic masking in that the mask token changes during each training epochs. Results showed that their transformer-based model (QWK= 0.80) outperformed the accuracy of the AES systems (0.66<QWK<0.77) that were based on traditional language models.

Rodriguez et al. (2019) compared the accuracy of BERT, XLNet, Bag of words (BoW) embedding (e.g., Word2Vec), and RNN models in AES on the ASAP dataset. Like BERT, XLNet is a transformer-based language model containing 110 million parameters that is trained bidirectionally. Unlike BERT that is trained on MLM and NSP tasks, XLNet used permutation-based modeling for training. In permutation-based modeling all the tokens are predicted in random order. This contrasts with MLM where only the masked tokens (15% of all the tokens) are predicted. Results of their study indicated that, compared to the BoW methods, the RNN models were more accurate, with a 6 % higher QWK, on average. Moreover, the models

containing an individual network, such as LSTM and a transformer-based model (e.g., BERT or XLNet) outperformed humans (QWK= 0.75) by about 4%. Regarding the individual models, BERT, XLNet, and the LSTM obtained very similar average QWK across all the essay prompts. These results demonstrate that BERT AES model has the potential to significantly improve the scoring performance of our current AES systems.

BERT has also been used to address the problem of data sparsity in AES. BERT uses self-supervised learning algorithms to capture subtle nuances in unlabeled data that were first identified during the training phase using the labeled data (e.g., Xue et al., 2021). This acquired knowledge inherent to the pre-trained BERT can then be "transferred" to the downstream task. In the context of transfer learning, the downstream task for AES is text classification. BERT has produced promising results in transfer learning on AES classification tasks (Sun et al., 2019). The approach described by Sun et al. (2019) was to use BERT as a pre-trained model on a large network with a large amount of unlabeled data, and further pre-train the general features of BERT on the self-supervised tasks (i.e., MLM and NSP) using the target data. Using transfer learning, the developed AES systems performed comparable or higher than (0.78<average QWK<0.79) the previous state-of-the-art deep learning AES systems (0.76<average QWK<0.76) with fewer training data (Cao et al., 2020; Dong et al., 2017; Hellman et al., 2019; Taghipour & Ng, 2016; Xue et al., 2021). More specifically, using the BERT encoder for essay representation on the ASAP dataset, Cao et al. showed that self-supervised learning algorithms are successful in transferring essay representations from one essay prompt to another without the need to train the model on all the prompts. That is, Cao et al. showed that using BERT to train essays in Prompt 5 of the ASAP data, a system can achieve a very high accuracy (QWK=0.80) on predicting the scores of the essays in the Prompt 6 of the data. This finding demonstrates that BERT as a self-

supervised learning model can solve the problem of data sparsity in the AES by transferring the learned essay representations from the labeled data on one essay prompt to the unlabeled data of another essay prompt.

Taken together, outcomes from the existing research demonstrate that BERT is a promising method for solving challenging AES problems. However, the vast majority of the applications conducted to-date used English language corpora, such as ASAP and TOFEL 11 (Ke & Ng, 2019). Very few studies have been conducted on languages other than English to develop AES systems using BERT. Only two studies were reported in the literature. Hirao et al. (2020) developed an AES system using a labeled dataset consists of 558 Japanese essays. They used the dataset to train three AES systems including BERT, LSTM, and machine learning (ML) with hand-crafted features (e.g., random forest). BERT achieved the highest accurate measure (QWK= 0.62) in essay scoring compared to the LSTM (QWK= 0.24) and the ML models (QWK= 0.52).

Li and Dia (2020) developed an AES system using a labeled dataset consists of 300 Chinese essays. They used a BERT network to obtain the sentence vectors for essays and then used a bidirectional LSTM (Bi-LSTM) network with two layers to extract the essay vectors. Because of data sparsity, the authors could not access a real training set that was written by Chinese students and scored by human raters, so they simulated 300 essays in Chinese for training their model. The model in which the BERT was combined with the Bi-LSTM outperformed (QWK= 0.60) the other models including the Bi-LSTM (QWK= 0.55) and the ML with hand-crafted features (QWK= 0.51).

These two studies provide evidence that BERT-based AES systems can yield improvements over the other neural network models, such as LSTM, for scoring essays written in languages

other than English. However, the reported reliability in these two studies is low compared to those studies conducted on English language essays. One reason for the comparatively low reliability may be attributed to the size of the datasets used in these two multilingual AES studies. Given the large number of parameters required in BERT, a small dataset used for tuning can easily be over fit leading to results with low accuracy (Ezen-Can, 2020). Hence, the purpose of this study is to expand the literature on multilingual AES by developing the first AES system for the Persian language using BERT. This study serves as a noteworthy improvement over the two multilingual AES studies currently available in the literature by training the system on a sufficient number of essays. We also describe and implement a unique model architecture that can be used by future researchers and practitioners to model multilingual AES data using BERT.

Chapter Summary

AES is a method where a machine attempts to provide scoring decisions by emulating how the written responses were scored by human raters. This method is popular because of the rapidly expanding use of online learning courses where it can be applied to efficiently and economically score students written tasks (e.g., Ramesh & Sanampudi, 2021; Uto, 2021; Wind et al., 2018). In addition, the rapid development of neural network models and NLP techniques has increased the accuracy of AES methods to the point where the reliability of the AES system is comparable to or even better than the scores provided by human raters (e.g., Beseiso et al., 2021; Shin & Gierl, 2020; Taghipour & Ng, 2016). More specifically, the development of transformer-based language models, such as BERT, have increased the accuracy of AES systems because these models consider the context in the text representation (Devin et al., 2018). Research has demonstrated that AES systems developed using a transformer-based model along with neural network layers can be used to outperformed human raters reliability by about 4%.

Almost all of the published AES studies have focused on essays written exclusively in English (Ramesh & Sanampudi, 2021). The studies on multilingual AES have been conducted on just a small number of languages (e.g., Hirao et al., 2020). Multilingual AES is an important research area because the language of assessment for many students throughout the world is not English.

Persian is an Indo-European language which is spoken by more than 110 million people around the world. To date, no AES system has been developed for this language. Persian is different from English orthographically and linguistically. Unlike the Latin script, Persian is written from right to left. Persian is also a cursive script in which alphabetic letters are attached to create a word. The morphological system and sentence structure in Persian are different from English. Persian has a wide variety of derivational and functional suffixes. Regarding the sentence structure, Persian follows subject, object, and verb structure while English does subject, verb, and object. Given these unique features of Persian language, the existing AES systems that were trained on other languages are not applicable to Persian. Thus, it is desirable to develop a unique AES system for the Persian language because Persian is the language of instruction and assessment at schools and higher education institutions throughout the world.

CHAPTER 3

METHOD

Dataset

The data used in my research contained 2,000 Persian essays. The essays were written by non-native Persian language learners from the Saddi Foundation, which is a Persian language education centre in Iran. The learners were female and male students of diverse nationalities located in the Middle East, North Africa, South Asia, East Asia, North America, and Europe. The essays were scored (i.e., labeled) holistically by the course instructors (i.e., raters). The raters were trained to evaluate the overall quality of the essays based on a holistic rubric (Urquhart & McIver, 2005) and assign a single score to each essay in a range of 1 (Elementary) to 5 (Advanced) (see Table 1). To reduce the subjectivity of holistic scoring, each essay was scored by two raters, and the average of the scores was reported by the Saddi Foundation. Therefore, the labels used in this study are the average of the two scores for each essay.

Table 1.

Descriptive Statistics for the Persian Essays

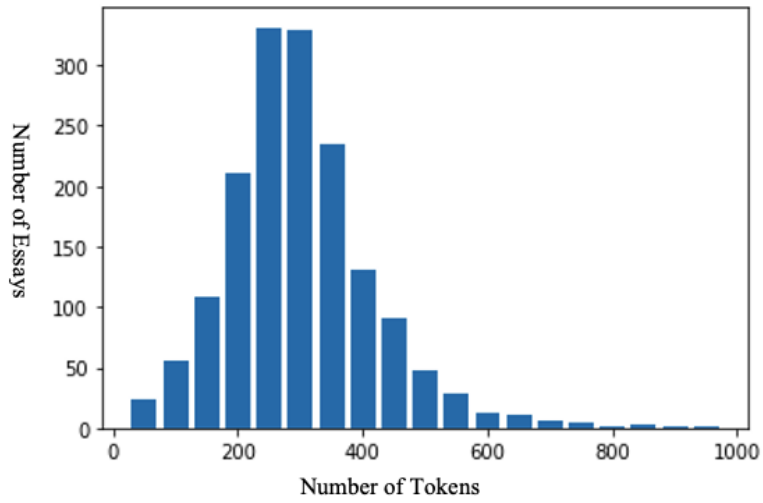
Level	Number	Average Length	Score
Elementary	538	124	1
Pre-Intermediate	476	143	2
Intermediate	387	153	3
Upper-Intermediate	361	192	4
Advanced	238	205	5
Total	2000	164	

The holistic rubric covered general writing criteria, including the variety and appropriateness of the vocabulary used, the variation and correct use of the tense of the sentences (e.g., the use of active and passive sentences, subject/verb agreement), sentence complexity (e.g., the correct use

of simple and complex sentences), mechanics of writing (punctuations and spelling), and the relevance and logical coherence of the sentences to the given prompt (i.e., the topic that was asked to write about). The essay prompts required students to describe their ideas about different real-life situations. For example, “describe your student life experience at the dormitory,” “how children should spend their free time,” and “describe your idea about eating at home or eating out” (see Appendix A and B). Across all of the essays in the dataset, the average length was about 164 words (see Figure 2).

Figure 2.

Essay Length Distribution

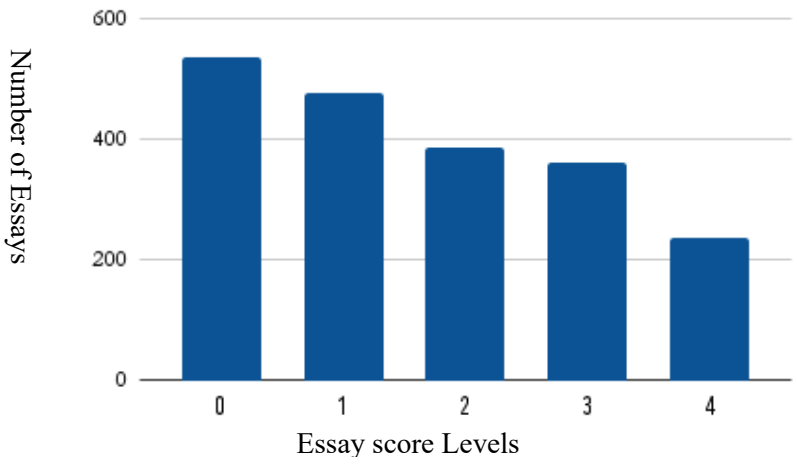


The essays were assigned a score from 1 to 5, corresponding to the proficiency levels described as Elementary, Pre-Intermediate, Intermediate, Upper-Intermediate, and Advanced. Students at the Elementary level had little or no exposure to Persian and had been learning basic reading, writing, and speaking skills. They were able to use memorized words and phrases and express factual information by manipulating grammatical structure. Learners at the Intermediate or Upper-Intermediate levels had an understanding of the language to successfully express a wide range of relationships, such as temporal, sequential, cause and effect with a wider range of

vocabulary use. Learners at the Advanced levels had sufficient mastery of the language to shape their writing skills to address different purposes and to clearly defend or justify a particular point of view. The number of students at each level was not equal. As Figure 3 shows, the highest number of essays were written by students at the Elementary level of language proficiency and the lowest number of essays were written by students at the Advanced level.

Figure 3.

Score Distribution of the Essays



Note. Essay scores from 0 to 4 are assigned to performance categories, using the following labeling system: 0=Elementary, 1= Pre-Intermediate, 2= Intermediate, 3= Upper-Intermediate, 4= Advanced.

Data Analysis

The data analyses consisted of text pre-processing, hyperparameter tuning, and model development. In text pre-processing, the essays were tokenized at both word and sub-word levels to interpolate between word-based and character-based tokenization using the Tensorflow text package in Python (Abadi et al., 2015). The name entities in the texts were extracted to enhance

the accuracy of the system (Oliveira et al., 2019). Then, transformer-based special tokens were added to make each sentence's beginning and ending understandable to the transformer system. Furthermore, in processing sequence data, it is prevalent for each input sentence to have different lengths. However, the input data for a deep learning model must have tensors with a specific shape. Therefore, padding and truncation strategies were used to make the shape of the sentences uniform to the tensor by adding zero values to sentences with fewer tokens (Lagouvardos et al., 2020). As the final stage of pre-processing, the training and testing datasets were randomly split, where 80 percent of the dataset was put aside for training, and the remaining 20 percent was used for evaluating the trained model in the testing phase (Berrar, 2019).

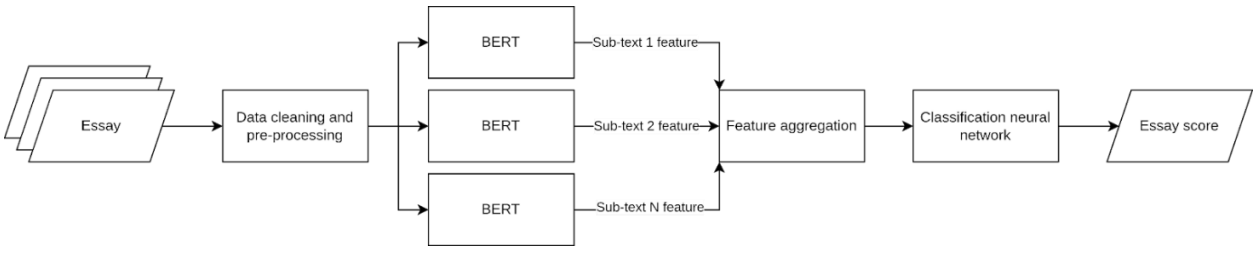
Model Architecture

Main Model: BERT-Based Model

The proposed transformer neural network model created for this research is presented in Figure 4. The proposed model consists of BERT, a recurrent neural layer, and a classifier. For all analyses, training will be performed on Google Colab Pro Cloud servers with 32 GB of RAM and an Nvidia Tesla P100 GPGPU.

Figure 4.

Model Architecture for Persian AES System

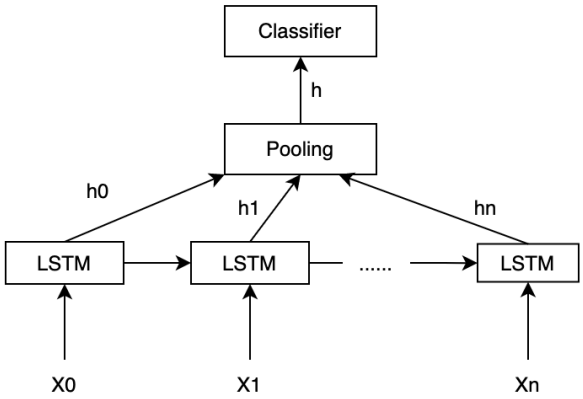


After pre-processing the data, the sequence of inputs was fed into the pre-trained Multilingual BERT system as shown in Figure 4. Depending on each essay's length, various layers were used for feature extraction. BERT generates embeddings for the words in a predefined window over the text. The predefined window length in this study is 256. This number is determined based on the distribution of essay length in the proposed dataset. Choosing the mean length of essays in the dataset ensures that a large number of inputs (i.e., words and sentences in a text) are fully captured by the model while limiting computational requirements, which would extend processing time if larger models were used.

Then, a recurrent network was used to aggregate the multiple embeddings into a single embedding after the transformer window swipes over the long text. The LSTM model was used as the recurrent layer to aggregate the BERT output at each time step. This model combines the data into a sequence of vectors having the same length relative to their temporal position and temporal dependency with respect to the features in the essays (Figure 5). LSTM was selected over other potential operations because the LSTM layers tend to produce more accurate modelling results of deep connections between sequential features that can be used to improve score prediction for AES applications (Qin et al., 2019). Taken together, the aggregated 768-dimensional embedding captured the general information of each embedding that required for the specific AES task in this study.

Figure 5.

Aggregation Layer for Multilingual BERT Model



Then, a classification neural network was used to classify texts based on their representations in the 768-dimensional space. The classification neural network is a single-layer, fully connected neural network used to map the feature space into the essay score space. Using the sigmoid activation in this layer converts the network output into a probability distribution over each essay score. For each input text, the selected essay score is the score level where the score level can be considered a specific class—in this study, there were five classes that range from 1-Elementary to 5-Advanced—with the highest probability in the neural network output.

In the last step, the combination of the BERT AES model and the following recurrent and fully connected layers were trained end-to-end on the essay scoring dataset. Fine-tuning the base pre-trained BERT AES model optimized the features produced by the model to be more suitable for the essay scoring task in this study. The feature mapping learned by the BERT AES model in this step clustered the essays with similar scores into the feature space.

Baseline Model: Word Embedding-Based Model

I also developed a word embedding-based model as a baseline against which the result of the proposed BERT-based model was be evaluated. Word embedding is a technique for

representing a word as a real number vector while preserving its meanings, semantic relationships, and alternative meanings using the distances among the words. Words that are closely related (i.e., similar meanings, similar locations in a sentence) should be located in a close vector space, while words that are far apart should be more distant from one another. For example, ‘location’ and ‘destination’ are semantically related words, so the reasonable embedding space would represent them as vectors that are not far apart. Word embedding has been critical to improving the performance of various natural language processing tasks, such as syntactic parsing and sentiment analysis (Socher et al., 2013). Word embedding is also an essential procedure in AES as many machine learning algorithms and almost all deep learning algorithms are incapable of processing strings or plain text in their raw form. Rather, they require text to be processed and represented as numeric input to perform accurate predictions.

Word2vec (Mikolov et al., 2011) is a commonly used word embedding technique. In fact, it is often considered the standard for pre-trained word embedding in text analysis (Mayfield & Black, 2020). Word2vec attempts to learn a geometrical representation of words from their co-occurrence information using predictive models. More specifically, Word2vec uses the two predictive models—which are the continuous bag-of-words (CBOW) and continuous skip-gram (CSG)—to compute the probability of a target word occurring in a particular context that is defined by neighboring words. While the CBOW model learns the embedding by predicting the current word based on its context, the CSG model learns by predicting the surrounding words given a current word. Using the two methods, Word2vec learns the relationships between the target word and context word one by one as it moves through every target word in a corpus. As a result, learning can take a significant amount of time when a text includes many words.

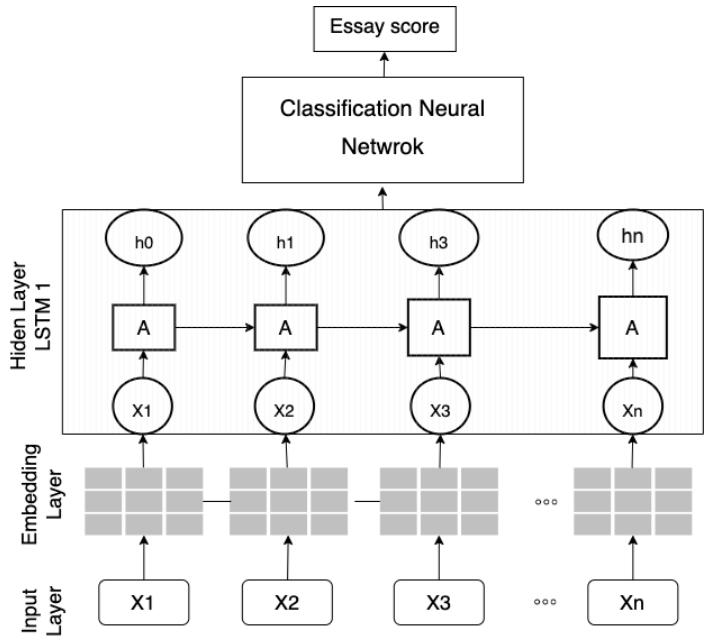
There are two versions of Word2Vec: Continuous Bag of Words (CBOW) and Continous Skip-Gram (CSG). The CBOW model learns the embedding by predicting the probability of each word based on its surrounding words or context constrained to the window size specified before model training. The CSG model learns by predicting the probability of the context of a word-gram neighboring words based on a set of words. Word2Vec tries to capture the co-occurrence of words one window at a time.

Some studies have demonstrated that Word2vec can learn vector representation more efficiently regardless of the language type compared to other word embedding techniques, such as GloVe (e.g., Naili et al., 2017; Berardi et al., 2015). The Word2Vec was selected as the baseline model in this study because it was the standard technique for text representation before the emergence of the transformer models (Mayfield & Black, 2020).

In order to not add to the complexity of the baseline model, the architecture follows that of the BERT-based model except for the text representation layer. Hence, instead of using the BERT transformer for the input layer of the model, I used the Word2Vec embedded matrix. The architecture for the baseline model is presented in Figure 6.

Figure 6.

Baseline Word2Vec Embedding Model



The model's input was the word embedded matrix and the hidden layer was an LSTM layer to learn the features end-to-end. A neural network was used as the final layer upon which essay classification was based.

Improving the Model Performance at Each Level Using Text Augmentation

The descriptive statistics in Table 1 reveal that the number of essays at each level are not equal. The number of essays at the Elementary level is the highest (n=538) while at the Advanced level it is the lowest (n=238). To develop an accurate AES system, a relatively large amount of data is required to train the model. Although models like BERT can be used to accurately predict essay scores with a relatively small dataset, student scores at the Advanced level are still needed to fine tune the BERT model. Therefore, we introduced and implemented a BERT-based data augmentation technique to increase the number of essays in the current study.

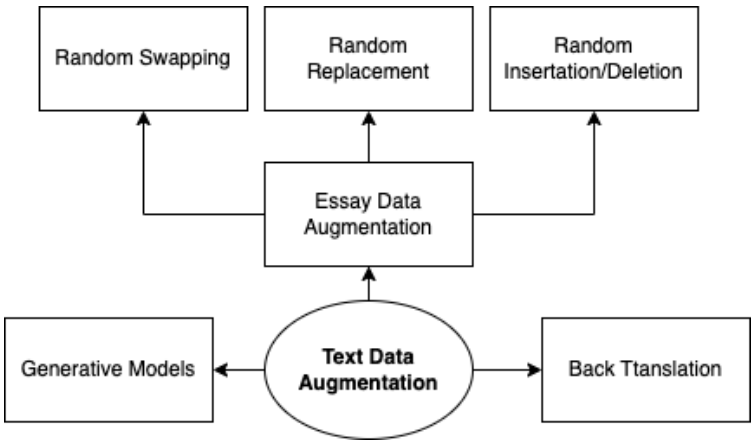
Text Data Augmentation Overview

Data augmentation is a process where the training data size can be artificially increased by generating different versions of real datasets without either collecting new data or annotating a large amount of existing training data. Data augmentation, as an analysis strategy, has been used in the fields of computer vision (e.g., Simard et al.,1998; Szegedy et al., 2014; Krizhevsky et al., 2017; Jiang et al., 2020) and speech (Cui et al., 2015; Ko et al., 2015) to help create more robust models, particularly when using smaller datasets. However, because it is challenging to come up with generalized rules for language transformation, universal data augmentation techniques to deal with data scarcity and insufficient data diversity in NLP are still novel, not well established, and rarely used in practice (Wei & Zou, 2019).

The most recent text data augmentation techniques are focused on using predictive language models (e.g., Kobayashi, 2018; Yu et al., 2018), word embedding models, and transformer-based models to generate data for different components (i.e., character, word, and sentence) of a text. To implement this technique, text is generated for a given sentence in a training set using different techniques, including essay data augmentation (e.g., Wei & Zou, 2019), back translation (e.g., Hayash et al., 2018), and generative modeling (Li et al., 2018). Figure 7 shows the text data augmentation models.

Figure 7.

Text Data Augmentation Models



Common essay data augmentation methods include Synonym Replacement (SR), Random Insertion (RI), Random Swap (RS), and Random Deletion (RD). With SR, *n* number of words are chosen randomly from the sentence and are then replaced with one of their synonyms chosen at random. For example, given the sentence "The focus of this dissertation is on using modern language processing techniques.", the SR method randomly selects *n* words (e.g., three), 'dissertation,' 'modern,' and 'techniques' and replaces them with 'thesis,' 'novel,' 'strategies,' respectively. RI finds a random synonym of a random word in the sentence and inserts that synonym into a random position in the sentence. This operation is done *n* number of times. For example, the synonym 'strategies' is added to a random position in the previous example: "The focus of this dissertation is on using modern *strategies* language processing techniques." The third technique, RS, randomly selects two words in a sentence and then changes their positions. This operation is done *n* number of times. For example, the position of 'modern' and 'dissertation' is randomly changed in the sentence "the focus of this *modern* is on using *dissertation* language processing techniques." With RD, each word with a given probability (*p)*

is randomly removed from a sentence. For example, the original sentence is shortened to “The focus of this dissertation is on using modern language processing”.

Although SR, RI, RS, and RD are common essay data augmentation methods, the *quality* of the augmented data is largely affected by the method that is used to perform the replacement (Wei & Zou, 2019). The most popular replacement approaches for augmenting text include a lexical-based replacement, word embedding-based replacement, and contextual bidirectional embedding replacement (Bayer et al., 2022).

Lexical-based replacement uses large-scale electronic lexical databases, such as WordNet (miller, 1995), that are developed based on the principles of lexical semantics. Lexical semantics is the relationship between lexical items, the meaning of sentences, and the syntax of sentences. With this approach, random words containing different parts of speech (e.g., noun, verb, adjective, adverb) that are not part of a named identity are substituted into the text based on specific probabilistic outcomes (e.g., Kobayashi, 2018), geometric distributions (e.g., Marivate & Sefara, 2020), Chi-square statistics (e.g., Wang et al., 2019), or maximize loss functions (e.g., Jungiewicz & Smywiński-Pohl, 2019). Although lexical-based replacement can be used to generate useful data, it is conducted at the word level, which means that this approach can result in the loss of semantic information for entire sentences. Moreover, lexical-based replacement is not beneficial for low-resource languages (e.g., Persian) which lack large lexical databases, such as WordNet.

Comparable to synonym substitution using a lexical database, embedding replacement uses pre-trained word embedding like GloVe, Word2Vec, or fastText to search for words with the nearest word vector from the embedding space that provides adequate fit, meaning that the context of the added text is not altered appreciably. To achieve this outcome, the words that will

be replaced are first translated into a latent representation space, and then words that have similar contexts are selected as replacements. The latent spaces are produced from a distributional hypothesis of distributional semantics (Firth, 1962), which is currently the most popular approach for embedding models (Bayer et al., 2022).

Embedding replacement approaches, such as Word2Vec and Glove, have performed well in studies where text augmentation improves the accuracy of models for text classification tasks (e.g., Rizos et al., 2019; Wang & Yang, 2015). The major limitation of embedding replacement for text augmentation is that it does not guarantee the preservation of the contextual meaning of each replacement. This limitation, in turn, can lead to distortions of the meaning of the text. For example, the word ‘open’ has different polysemy (e.g., ‘available,’ ‘honest,’ or ‘not covered’) that should be selected based on the context of the word within a sentence.

Fortunately, transformers can solve the problem of understanding context, as initially described with embedding replacement, by predicting subsequent or missing words based on the previous or surrounding context of each word. Hence, contextual bidirectional embedding like BERT can filter out words that do not fit and, instead, identify more appropriate words leading to more predictive outcomes because the vector representation is more accurate. Given that the transformer-based models encode longer text sequences, they are contextually “aware” of the surrounding words, resulting in a more accurate text augmentation. The contextual understanding of BERT for text augmentation is achieved though masking specific tokens in a text and predicting those tokens using the training set. The text augmentation algorithm is shown below (Marivate & Sefara, 2020):

Input: Training Dataset D_{train}

Pertained model $G \in \{encoder, seq2seq\}$

1. Fine-tune G using D_{train} to obtain G_{tuned}
2. $D_{synthetic} \leftarrow \{\}$
3. **foreach** $\{x_i, y_i\} \in D_{train}$ **do**

   Synthesize s examples $\{\hat{x_i}, \hat{y_i}\}_p^1$ using

   G_{tuned}

   $D_{synthetic} \leftarrow D_{synthetic} \cup \{\hat{x_i}, \hat{y_i}\}_p^1$
4. **end**

Given a training dataset $D_{train} = \{x_i, y_i\}_n^1$, where $x_i = \{w_j\}_m^1$ is a sequence of m words, y_i is the associated label, and a pre-trained model G, we want to generate a dataset of $D_{synthetic}$. This algorithm describes the data generation process. For all augmentation methods, we generate a S = 1 synthetic example for every example in D_{train}. Thus, the augmented data is the same size as the original dataset. For example, if the size of an essay x_i is n number of tokens, then the augmented essay $\hat{x_i}$ also contains the same n number of tokens. Moreover, given the use of training data D_{train} in the process, the contextual word meanings are included in the augmenting new data $D_{synthetic}$.

Implementing Text Augmentation in AES

Transformers have been used for data augmentation in various text classification tasks such as sentiment analysis (e.g., Abonizio et al., 2021; Feng et al., 2022) and short-answer grading (e.g., Lun et al., 2020; Wu et al., 2019) to solve data sparsity problems using supervised deep learning models. However, there is just one example of the application of transformers for data augmentation in AES using multilingual text. Jong et al. (2022) used back translation to add essays to the ASAP dataset. In back translation, the original data is translated into another language and then translated back to the original language in order to obtain new data in the

original language (Hayashi et al., 2018). Using this approach, the entire text is rewritten without replacing individual words. Jong et al (2022) used Google's Cloud Translation API service to translate the English essays into French and perform back translation from French into English. They also used the Baidu Translation API service to back translate some of the essay prompts into Chinese. After obtaining the back-translation essays, the score of the original essays were adjusted for the augmented essays by providing a condition that is based on the length of the augmented essays. For example, if p is the condition for judging whether an essay has a length greater than 300, when $p(\hat{x_i}) = 1$, it shows that the length of essay $\hat{x_i}$ is greater than 300, so the maximum, minimum, or mean of the most frequent scores of the essays $(x_i)_n^1$ in prompt (i) with the same length is assigned to the augmented essay $\hat{x_i}$. Using this approach, Jong et al. (2022) improved the predictive performance of their AES model by 0.2%, on average, on five of the 8 prompts (i.e., 2, 3, 4, 6, and 8) using the ASAP dataset, which was first described in Chapter 2.

The study by Jong et al. (2022) indicated that text augmentation methods could address the problem of data sparsity in AES. However, there is no guarantee that the difficulty level of texts remains the same after translation (Hale & Campbell, 2002). Considering the importance of score reliability in text augmentation, which could be altered if the difficulty of the text is changed, Kumar et al. (2021) proposed a conditional data augmentation method using a pre-trained BERT AES model for sentiment classification purposes. They utilized BERT's and GPT2 segment embeddings to condition their model on the labels of the original data. Accordingly, they conditioned the models by pre-pending labels to all examples of a given score level. Then, they trained the model with different insertion training and masking given by two different methods:

- BERT $_{word}$: Replace a word w_i in text x_i with a mask token $< mask >$ in a random text $(xi)\,{}_1^n$ in a same score level or
- BERT $_{span}$: Replace a continuous chunk of k words $w_i, w_{i+1} \ldots w_{i+k}$ with a single mask token $< mask >$.

Result of the Kumar et al. (2021) study showed that the BERT AES model performed better than the GPT2 in predicting the masked tokens. Further, the BERT AES model provided a good balance between diversity and semantic fidelity. In addition, the diversity of the generated data can be controlled by varying the masking ratio.

Although Kumar et al.'s (2021) study demonstrated the importance of BERT in text augmentation for classification tasks, there is no specific study in the AES literature that matches the difficulty level of the generated texts with the original data. Given the data sparsity issue in multilingual AES, the augmentation method can be a solution to solve this problem. Hence, the method described by Kumar et al. (2021) will be implemented in the current study to address the data sparsity problem in the Persian dataset.

To implement this augmentation method, the results of the current study were expanded by using the pre-trained BERT to augment 20% of the original data at each score level using BERT and the conditional method proposed by Kumar et al. (2021). Thirty percent of the tokens were masked and 30% of the token chunks (window size 6) in each text for the Elementary, Pre-Intermediate, Intermediate, and Upper-Intermediate levels. Because the text length in the Advanced level was more than the length of the texts in other score levels, 45% of the tokens and 40% of the token chunks of each text at this score level were masked. BERT was trained to replace the masked tokens and chunks with the most similar tokens and chunks in texts with the

same score level. By replacing parts of tokens and sentences in a text, new data were created to augment the training and testing datasets.

Performance Metrics

Model validation in AES depends on comparing the similarity between the model performance and the human raters (Attali, 2013; Chung & Baker, 2003; Williamson et al., 2012). In this comparison, human judges are considered the 'gold standard' and function as the explicit criterion for evaluating the performance of the AES system (Latifi, 2016). Various validity coefficients have been adopted as evaluation metrics in previous studies to measure agreement. These measures include the Quadratic Weighted Kappa (QWK), the Kappa coefficient, and error analysis. QWK and Kappa are often used together as consistency measures in AES (Shermis, 2014). However, the most widely used performance evaluation measure for AES is the QWK (Williamson et al., 2012).

QWK is implemented when ordered categorical data is used. Where i represents a human-rated score, j represents a machine-rated score, and N is the number of possible ratings, a weight matrix W can be constructed as follows:

$$W_{i,j} = \frac{(i-j)^2}{(N-1)^2}. \tag{8}$$

Then a matrix O is constructed so that $O_{i,j}$ represents the number of essays that receive a rating i by the human and a rating j by the machine. An expected count matrix E is computed as the outer product of histogram vectors of the two ratings. The matrix is normalized so that the sum of elements in E and O are the same. QWK can then be calculated as:

$$QWK = 1 - \frac{\Sigma_{i,j} W_{i,j}\, O_{i,j}}{\Sigma_{i,j} W_{i,j}\, E_{i,j}}. \quad (9)$$

QWK varies from 0 (random agreement between raters) to 1 (complete agreement between raters). Typically, values between 0.60 and 0.80 QWK are used as a lower bound estimate for an acceptable reliability outcome using human raters in a high-stakes testing situation (Williamson et al., 2012).

The Kappa score provides a chance-corrected index and is computed based on the ratio of the proportion of times the agreement is observed to the maximum proportion of times that the agreement is made while correcting for chance agreement (Siegel & Castellen, 1988). Equation 10 shows the probability of interrater agreement where p_0 is the observed agreement among raters and p_e is the theoretical probability of chance agreement:

$$k = 1 - \frac{1-p_0}{1-p_e}. \quad (10)$$

Kappa ranges from 1, when the agreement is perfect, to 0 when the agreement is not significantly better than chance. Landis and Koch (1977) proposed values for interpreting Kappa. They claimed a k value < 0 indicates less than chance agreement; k values of 0.01–0.20 represent slight agreement; k values of 0.21–0.40 indicate fair agreement; k values of 0.41–0.60 represent moderate agreement; k values of 0.61–0.80 represent substantial agreement, and k values of 0.81–0.99 indicate almost perfect agreement. Thus, by convention, a Kappa value greater than 0.80 is considered excellent agreement and a value greater than 0.60 is considered good agreement.

In addition to score consistency, machine performance can also be evaluated by calculating score prediction accuracy. Accuracy is the number of classifications a model correctly predicts

divided by the total number of predictions made. To understand the performance of the model at each score level, the error analysis of the model is evaluated using the confusion matrix (see Table 2). The confusion matrix helps to understand the misclassification rate or error rate when more than two scale points are classified (Visa et al., 2011). The matrix contains a measure of precision and recall as well as the F-score index.

Table 2.

Confusion Matrix Cell Definitions

		Predicted Score	
True Score			
	True Positive (TP)	False Negative (FN)	
	False Positive (FP)	True Negative (TN)	

Precision, as presented in Equation 11, refers to the number of true positives divided by the total number of positive predictions. It quantifies the number of positive score level predictions that actually belong to the positive score level.

$$Precision = \frac{TP}{TP + FP} \tag{11}$$

Recall, shown in Equation 12, refers to the number of true positives divided by the sum of true positives and false negatives. It quantifies the number of positive score level predictions made out of all positive examples in the dataset.

$$Recall = \frac{TP}{TP + FN} \tag{12}$$

F-score (see Equation 13) combines the precision and recall of a model and it is defined as the harmonic mean of the model's precision and recall outcome (Visa et al., 2011).

$$F_{score} = \frac{2 * Precision * Recall}{Precision + Recall} \quad (13)$$

Chapter Summary

The purpose of this dissertation was to develop the first AES system for the Persian language using multilingual BERT. This study also provides the first architecture for an AES system that can be used in the Persian language. My research contributes to the field of educational testing by building on the multilingual AES literature. The results of this study will help demonstrate the application of BERT on a language with a completely different script system than English. BERT is a transformer-based language model that is designed to pre-train deep bidirectional representations by jointly conditioning on both the left and right context in all layers in a self-supervised fashion. This means it was pre-trained on millions of unlabeled—meaning data with no human annotation—publicly available texts using MLM and NSP techniques. The pre-trained BERT AES model can be fine-tuned with an additional output layer to create state-of-the-art models for a wide range of text classification tasks. BERT achieved state-of-the-art performance by significantly improving on other RNN language models such as GloVe and Word2Vec (Mikolov et al., 2013) in previous text analyses (Delvin et al., 2018).

In addition, the conditional data augmentation method described by Kumar et al. (2021) was used to address the data sparsity problem in the Persian dataset. To implement this augmentation method, new data were added to the original data at each score level. Thirty percent of the tokens were masked and 30% of the token chunks (window size 6) in each text were produced for the Elementary, Pre-Intermediate, Intermediate, and Upper-Intermediate levels. BERT was trained to replace the masked tokens and chunks with the most similar tokens and chunks in texts with the

same score level. By replacing parts of tokens and sentences in a text, new data was created to augment the training and testing datasets.

Multiple performance measures were used to evaluate agreement between the AES system and the human raters. These measures include the Quadratic Weighted Kappa (QWK), the Kappa coefficient, and error analysis. QWK and Kappa are used almost universally as consistency measures in AES. QWK is a weighted Kappa score used to overcome the problem inherent to Kappa, which is that it does not account for the degree of disagree. In addition to score consistency, performance can also be measured with an error analysis to evaluate score prediction accuracy. Accuracy is the number of classifications a model correctly predicts divided by the total number of predictions made. To understand the performance of the model at each score level, the error analysis of the model is evaluated using the confusion matrix. The confusion matrix contains a measure of precision and recall as well as the F-score index.

CHAPTER 4

RESULTS

This chapter is organized into five sections. The first section is an overview of the processes conducted to obtain and implement the data for the current study, including ethics requirements and the data pre-processing techniques. The second section describes the implementation of the AES architectures along with the statistical performance results. The performance of the AES models in each score level is also interpreted. The third section describes a technique—text augmentation—that improved the BERT AES model results presented in section two. The fourth section presents the result of the developed model using text augmentation. The fifth section contains a succinct summary of the chapter.

Section One: Ethics Review, Data Collection, Data Pre-Processing, and Text Pre-Processing

Ethics approval was received from the Research Ethics Board 2 (REB2) at the University of Alberta Research Ethics Office (see Appendix C) The REB committee reviewed the process of my data collection, including participants' information and the design of the current study. Given that the data used in this study were collected previously and released by the Saddi Foundation, it was considered secondary data analysis. There were three requirements that the data satisfied before receiving approval from the REB committee: (a) the data must be de-identified before being released to the researchers, (b) outcomes of the analyses must not allow re-identifying participants, and (c) use of the data must not result in any damage or distress to human participants. The study was approved because these ethical requirements were satisfied.

Data pre-processing was the first step. The original data—2000 essays in Persian with their associated scores—were provided in the extensible markup language (xml) format consisting of

an element tree class of information about the essays. The data were imported and read in Google Colab Pro, and Python 2.10.0 modules were imported and used for data pre-processing. Data pre-processing requires serialization. Serialization is a way to convert xml data structure into a linear form that can be stored or transmitted over a network. Then, the data were organized into a list of essays and their corresponding scores using the Python module NumPy v1.23. The purpose of this module is to provide a multidimensional array object, various derived objects (e.g., masked arrays and matrices), and an assortment of routines for fast operations on arrays, including mathematical, logical, shape manipulation, sorting, selecting, basic linear algebra, basic statistical operations, and random simulation.

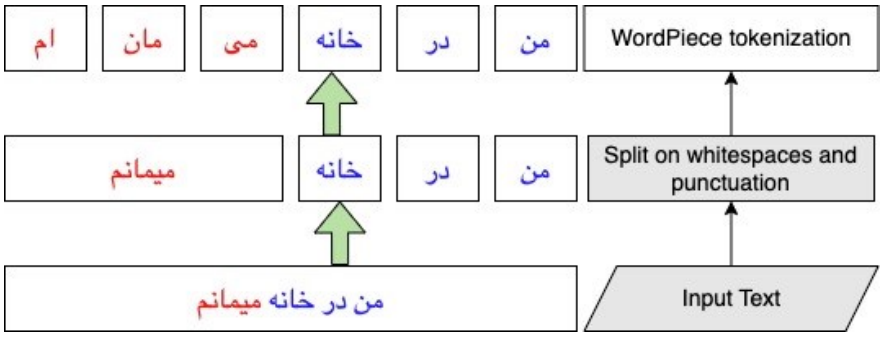
Next, text pre-processing was conducted. Text pre-processing is the end-to-end transformation of raw text into a model's integer inputs. Integer inputs help decrease the noise in the data so the model can learn more efficiently. Text pre-processing was initiated with tokenization using BertTokenizer in TensorFlow. Tokenization is a fundamental step in both traditional NLP methods like Count Vectorizer and deep learning-based architectures like transformers. Tokenization is used to separate a piece of text into smaller units. These smaller units, called tokens, can be either words, characters, or subwords (e.g., suffixes). BertTokenizer is a text splitter that applies an end-to-end text string to produce tokens. The WordPiece tokenization method in BERT operates as an intermediary between byte pair encoding (BPE) (Gage, 1994) and the Unigram Language Model (ULM) approaches. BPE is a data compression technique that iteratively replaces the most frequent pair of bytes in a sequence with a single unused byte (Sennrich et al., 2016). ULM is a statistical language model that analyzes the probability of one word sequence.

Once the ULM approach is implemented and each word is separated as a token in a text, WordPiece considers the probability of each word relative to any other words that come before it in the text. WorkPiece works by splitting words either into the full forms (e.g., one word becomes one token) or into word pieces—where one word can be broken into multiple tokens. For example, using the ULM technique, the sentence “I go surfboarding” is first tokenized into three words as ‘I’ ‘go’ ‘surfboarding’. Then, WorkPiece splits the ‘surfboarding’ into multiple tokens such as ‘surf’ ‘##board’, ‘##ing’. BERT uses specific tokens such as [##], [CLS], and [SEP] which are used as a prefix for word relation tokenization. These prefixes serve as an indicator of the beginning of a sentence and the indicator of the end of a sentence, respectively.

Because Persian has a variety of prefixes and suffixes, we used ParsBERT (Farahani et al. 2020) for WordPiece tokenization. ParsBERT was trained on about 40 million Persian sentences using the MLM and NSP strategies in BERT. The resulting number of tokens in this study (after pre-processing) was 100,000, which consisted of content and functional words as well as BERT-specific tokens, such as [PAD], [UNK], [CLS], [MASK] [SEP], and [##]. Figure 8 shows an example of the tokenization process for the Persian language based on the WordPiece method.

Figure 8.

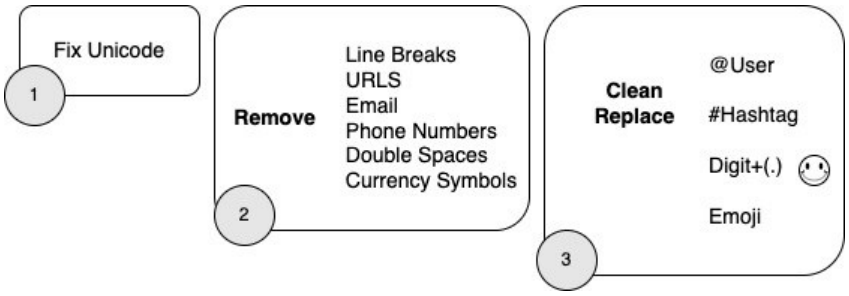
Subword Tokenization Using ParsBERT



After tokenization, the non-alphabetic words, numbers, characters (e.g., @, %…, ۱و۲ ☺), and name entities (e.g., name of cities and places) are eliminated from the text while punctuation is kept and treated as separate words. A complete representation of the text-cleaning process is shown in Figure 9.

Figure 9.

Text Cleaning Process



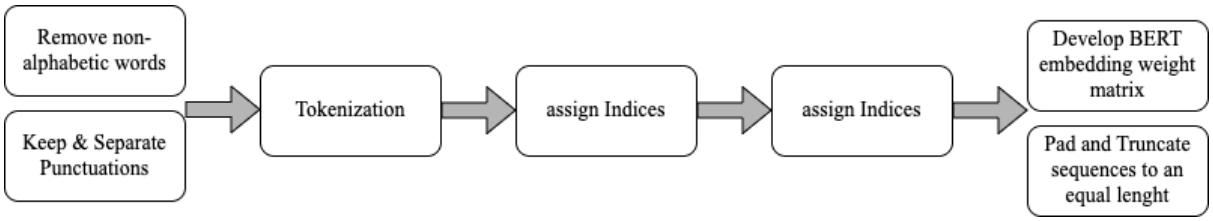
After text cleaning and tokenization, the text pre-processing step continues by transferring the tokens to the tensors. In this study, 40,250 tokens were transferred. Each token was then assigned a unique numeric index so that the index matched the location of the word in an embedding matrix.

The length of each text must be matched to the size of the tensors in BERT. In order to keep all the information in the texts, the highest size of the tensors (512) was applied to the texts. The mean of the number of tokens in texts was 450. Hence, BERT 512 covered all the information in the text. Given that every element in a specified tensor should contain a value, for texts with less or greater than 512 tokens, two different strategies—padding and truncation—were used to decrease and increase the number of tokens, respectively. The padding parameter was set to `True` to pad the shorter sequences in the batch in order to match the longest sequence, which is set to 512. For truncation, the sequences of input longer than 512 were truncated to a shorter

length using the `Slice` function. Then, the sliced arrays of an essay were added using the `Sum` function. All of the steps required for text pre-processing are summarized in Figure 10.

Figure 10.

Text Pre-Processing Steps



Word2Vec served as a baseline comparison in my study. Text pre-processing for the Word2Vec model was conducted using HAZM (please see https://github.com/sobhe/hazm). HAZM includes many pre-process tools such as normalizer, tokenizer, and part of speech labeler. This tool uses Persian word files, verb files, suffixes, and prefixes of verbs. The text cleaning operation replaces items such as links, emojis, emails, hashtags, and numbers with the correct formats, then divides the sentence according to the " " (space) character. The tokenized input was mapped to the pre-trained Persian Word2Vec model (Davoudi & Mirzaei, 2021) for text representation.

Section Two: Results of the Hyperparameter Tuning

Several neural network layers were developed for the AES models. I optimized the layers by examining different values for hyperparameters, including regularization parameters along with the activation functions, to produce the best results using grid search. Grid search is a technique in machine learning that is used to evaluate every possible hyperparameter—the training variables set manually with a pre-determined value—to select the best configuration for the learning weights in the model.

Previous research (e.g., Gang et al., 2020; Wang et al., 2016) demonstrated that RNN models, including the LSTM in which every step considers both the current input and the previous outputs, provide the best results when the text contains long dependencies. Therefore, I chose the LSTM as the aggregation layer and compared several different numbers of nodes, hidden layers, dropout, momentum, decay rate, and activation functions to produce the best results. In the current study, Stochastic Gradient Decent (SGD) was used as the optimization technique. Every possible hyperparameter configuration in Table 3 was considered using grid search. Table 3 depicts the candidates and the selected hyperparameters.

Table 3.

Hyperparameters for the BERT Model

Layer	Parameter Name	Candidate Values	Selected Value
BERT $_{multilingual}$	Number of parameters	110M	110M
	Transformer blocks	12 layers	12 layers
	Attention heads	12	12
	Hidden neurons	768	768
Dropout	Dropout rate	0.2-0.5	0.2
LSTM	Decay Rate	0.96-0.97	0.97
	Activation Function	relu, sigmoid	sigmoid
	Learning Rate	$0.1\text{-}10^{e\text{-}6}$	$10^{e\text{-}3}$
	Momentum	0.5-0.9	0.5
Dense	neurons	100	100
Model compile	Epoch	15-30	24
	Batch Size	128	128

After developing the BERT AES model architecture, to train and tune the model, the indexed tokens were mapped into the BERT-base multilingual uncased as the input layer. BERT-base multilingual uncased is a package that can be used with languages that do not have upper and

lower case texts. The uncased version was preferred over the cased version because there is no upper/lower case regulation in Persian. The BERT AES model contains 110,000,000 parameters, 12 transformer block layers, 768 hidden layers, and 12 attention heads. The dropout rate was set to 20% to avoid model overfitting while retaining model accuracy. The optimal decay rate to prevent over-flatting the updated weights was set to 0.97. The sigmoid activation function resulted in optimal model performance. The medium learning rate (10^{e-3}) led to higher model convergence. In addition, 0.5 was used as the optimal momentum rate to accumulate the gradient of the past steps. The aggregated weights from the LSTM layer were fed into a dense layer containing 100 neurons. Finally, the best-performing model was compiled at 24 epochs using 128 as the batch size.

For the baseline model using word embedding, I implemented the pre-trained Persian word2vec. The model is created using content from Common Crawl and Wikipedia. It is trained using CBOW with position weights, in dimension 300, with character n-grams of length 5, and a window of size 5 and 10. The model gives 300-dimensional vector outputs per token. The output vectors map words into a meaningful space where the distance between the vectors is related to the semantic similarity of words. For this study, the model was fined tuned using the hyperparameters presented in Table 4.

Table 4.

Hyperparameters for the Word Embedding Model

Word2Vec Variables	Initial Setting	Optimized Setting After Tuning
Number of dimensions	300	300
Window size	5	10
Minimum Word count	40	20
Learning rate	0.05	0.15
Epochs	10	10

As shown in Table 4, the number of dimensions was kept to 300, which can be considered a reasonably large value since it resulted in a high accuracy rate in previous studies (e.g., Dong et al., 2017). The minimum word count for training is the threshold for ignoring the words with fewer total frequencies in the context. Initially, the minimum word count was set to 40. Increasing the minimum word count to 50 (i.e., the highest possible value) resulted in a negligible increase in the model's accuracy. However, changing it to lower values like 5 (as an extreme case) led to high variance and overfitting, which resulted in a score that decreased from 0.74 to 0.64. In addition, changing the minimum word count to 20 raised QWK to 0.75. Therefore, the optimized minimum word count was set to 20. Another parameter in tuning Word2Vec is the window size. When the window size increases, more context can be captured for estimating the weights of words. However, large window sizes can also decrease the quality of model training (Levy & Goldberg, 2014). Thus, I carefully increased the window size of the initial setting from 5 to 10 to achieve an optimal result. The fined-tuned Word2Vec model was incorporated into the Bi-LSTM model with the same tuning parameters used in the BERT AES model.

Section Three: Results of the AES Models

Word Embedding-Based AES Model

The model consisting of a pre-trained Word2Vec and a Bi-LSTM produced a substantial agreement between the AES system and the human raters. As shown in Table 5, the model was highly reliable (QWK=0.75). The model's performance was the highest at the Elementary level (QWK=0.68) and the lowest at the Advanced level (QWK= 0.54). In addition, the κ coefficient (κ= 0.82) was above the accepted (κ>0.80) consistency level. κ also showed that the machine and human raters' score agreement is high (>0.80) for each score level except for Advanced (κ=0.67).

Table 5.

Word Embedding Model Performance on Each of the Score Levels

	Word2Vec+LSTM	
	QWK	Cohen's Kappa
Elementary	0.68	0.87
Pre-Intermediate	0.66	0.83
Intermediate	0.61	0.79
Upper-Intermediate	0.60	0.83
Advanced	0.54	0.71
Total	0.75	0.93

The Word2Vec and Bi-LSTM model resulted in about a 71% classification accuracy for the total number of essays scored by the system (see Table 6). Table 6 also demonstrates the performance of the system at each performance category.

Table 6.

Accuracy Scores by Level for the Word Embedding Model

Overall Accuracy		0.71	
Performance Level	Precision	Recall	F-Score
Elementary	0.73	0.86	0.78
Pre-Intermediate	0.69	0.81	0.73
Intermediate	0.65	0.79	0.71
Upper-Intermediate	0.67	0.82	0.74
Advanced	0.52	0.55	0.54

The Precision column in Table 6 shows that among all the essays correctly classified by the AES system, more than 65% in each performance level, except for Advanced, were scored the same by the human raters (i.e., true positive). The Recall column indicates that among all the essays that were wrongly classified, more than 75% of them in each score level except for Advanced were considered wrong classifications (i.e., true negative). The F-Score column shows the average for the precision and recall indices. The F-Score for the Elementary level is the highest (0.78) and the for the Advanced level the lowest (0.54).

BERT-Based AES Model

The Word2Vec and Bi-LSTM model served as the baseline in my research. The results for the BERT AES model are of primary interest. The proposed BERT AES model produced very accurate results, particularly when compared to the baseline. As shown in Table 7, the model consisting of BERT and LSTM performed with high reliability (QWK=0.84). The model's performance is the highest at the Elementary level (QWK=0.80) and the lowest at the Advanced level (QWK= 0.56). The κ reliability is above (κ= 0.93) the accepted (κ>0.80) consistency measure (Table 7). κ shows that the machine and human raters' scores are in high agreement with one another (> 0.80) at each score level, except for Advanced (κ=0.71).

Table 7.

BERT Model Performance on Each of the Levels

	BERT+LSTM	
	QWK Score	Cohen Kappa Score
Elementary	0.80	0.90
Pre-Intermediate	0.69	0.91
Intermediate	0.61	0.85
Upper-Intermediate	0.61	0.87
Advanced	0.56	0.71
Total	0.84	0.93

The result from the accuracy measures shows that about 73% of the total number of essays were correctly scored by the system (see Table 8). Table 8 also includes the performance of the system for each score level. The Precision column in Table 8 shows that among all the essays that were considered as correctly classified by the AES system, more than 70% in each score level except for Advanced were scored the same by the human raters (i.e., true positive).The Recall column indicates that among all the essays that were incorrectly classified, more than 70% of them in each score level—except for Advanced—were considered as incorrect classifications (i.e., true negative). In this study, the high precision and recall and their average (i.e., F Score) values indicate that the classification overlap of the model was fairly small and that, overall, the model performed with high accuracy. In addition, the F-Score values in Table 8 show that the system had the highest performance in scoring the essays at the Elementary level and the lowest performance in scoring of the Advanced level.

Table 8.

Accuracy Scores of the BERT Model at Each Level

Overall Accuracy		0.73	
Performance Level	Precision	Recall	F-Score
Elementary	0.82	0.86	0.84
Pre-Intermediate	0.78	0.75	0.76
Intermediate	0.70	0.72	0.71
Upper-Intermediate	0.73	0.70	0.71
Advanced	0.60	0.64	0.62
Average	0.72	0.73	0.72

Error analysis shows that each level had some overlap with the adjacent levels (Table 9). For example, the Elementary level had the highest overlap with the Pre-Intermediate level and vice versa. Also, the Upper-Intermediate and the Advanced levels had the highest number of overlaps with adjacent performance levels. However, when the levels are more distant, the number of classification overlaps was noticeably less. For example, there was zero overlap between the Elementary and Advanced, Elementary and Upper-Intermediate, and Pre-Intermediate and Advanced levels.

Table 9.

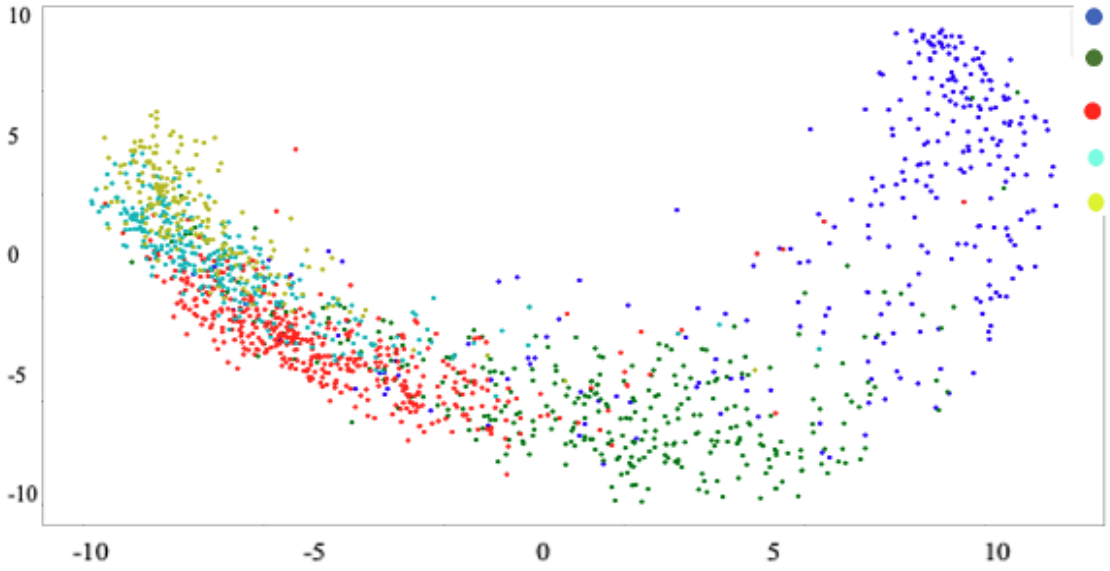
Confusion Matrix of the BERT Model at Each Level

Level	Elementary	Pre-Intermediate	Intermediate	Upper-Intermediate	Advanced
Elementary	**59**	5	4	1	0
Pre-Intermediate	10	**69**	8	3	2
Intermediate	2	13	**71**	8	5
Upper-Intermediate	0	3	15	**66**	12
Advanced	1	0	2	13	**28**

A visualization of the results in Table 9 is presented in Figure 11. Each circle represents an essay and the score level of the essays are shown with different colors. The representation of the essays in a high-dimensional space is transformed onto a two-dimensional space for the purpose of visualization. As a result, the values on the x- and y-axis are arbitrary. The values in Figure 11 are provided simply to illustrate the distance between the essays. The essays in each score level are clustered from Advanced to Elementary (upper left to upper right). As it is depicted in Figure 11, each score level has overlap with its adjacent level. In addition, the majority of the overlap occurs between the Advanced, Upper-Intermediate, and Intermediate levels.

Figure 11.

Representation of Essay Classification



Note. The levels of the essays are represented in color where Blue= Elementary, Green= Pre-Intermediate, Red= Intermediate, Light blue= Upper-Intermediate, Yellow= Advanced.

Model performance was the highest at the Elementary level and lowest at the Advanced level. The Advanced level yielded a QWK of 0.56 (model average 0.84), a Kappa of 0.71 (model average 0.93), Precision score of 0.63 (model average 0.72), a Recall score of 0.62 (model average 0.73), and a F-Score of 0.62 (model average 0.72). In addition, the results from the error analysis in the confusion matrix reveal that essays were misclassified in the Upper-Intermediate level with its adjacent levels most commonly. This result may be occurring because the descriptors used with the holistic rubric are too similar between score levels. When there are small differences between rating levels, the machine usually needs more or different types of data to understand the granularity and nuances between score levels that are similar to one another (Ormerod et al., 2021). At this point in the study, I can conclude that the Upper-Intermediate and the Advanced levels did not contain enough essay data for the AES system to

learn the distinction between the content in those two score levels. To evaluate this interpretation and to attempt to correct this problem, text augmentation was used.

Section Four: Improving the Model Performance Using Text Augmentation

Text Augmentation Results

Descriptive statistics of original and the augmented data are represented in Table 10. The amount of text generated increased the average length of the essays by about 20% across each performance level. However, the mean length of the augmented data in each level remained almost the same, 95% CI [10, 25], as that of the original data.

Table 10.

Descriptive Statistics of the Augmented Texts

Score Levels	Original Data		Augmented + Original Data	
	Number	Average Length	Number	Average Length
Elementary	538	124	644	132
Pre-Intermediate	476	143	572	135
Intermediate	387	153	456	165
Upper-Intermediate	361	192	434	202
Advanced	238	205	287	225

The results of our Persian AES model both before and after text augmentation is summarized in Table 11. The results showed that the performance-developed Persian AES (QWK= 84%, κ=88%) was improved after including the augmented texts in the original data (QWK=96%, κ=96%). Likewise, the performance (QWK) of the model at each score level was improved after text augmentation. The model performance was improved the most at the Advanced level (+0.35) and the least at the Elementary level (+0.16).

Table 11.

AES Model Performance in Each Score Level Using Text Augmentation

Score Levels	Pre-augmentation		Post-Augmentation	
	QWK	Kappa	QWK	Kappa
Elementary	0.80	0.89	0.96	0.98
Pre-Intermediate	0.69	0.85	0.97	0.93
Intermediate	0.61	0.76	0.83	0.81
Upper-Intermediate	0.61	0.77	0.86	0.84
Advanced	0.56	0.76	0.91	0.82
General	0.84	0.88	0.96	0.96

The results from the accuracy measures after text augmentation shows that 91% of the total number of the essays were correctly scored by the system (see Table 12). Table 12 also includes the performance of the system for each score level. The Precision column in Table 12 shows that among all the essays that were considered as correctly classified by the AES system, more than 90% in each level were scored the same by the human raters (i.e., true positive). The Recall column indicates that among all the essays that were incorrectly classified, more than 84% in each score level were considered as incorrect classifications (i.e., true negative). In this study, the high precision and recall and their average (i.e., F Score) values indicate that after augmenting the text, the BERT AES system could classify the essays in each score level with higher accuracy than without augmentation. In addition, the F-Score values in Table 12 show that the system had the highest performance in scoring the essays at the Elementary level and the lowest performance in scoring of the Intermediate level.

Table 12.

Accuracy Scores by Level After Text Augmentation

Overall Accuracy		91%	
Performance Levels	Precision	Recall	F-Score
Elementary	94	98	96%
Pre-Intermediate	94	93	94%
Intermediate	90	84	87%
Upper-Intermediate	89	88	88%
Advanced	90	89	90%
Average	91	90	90%

Error analysis was also conducted. The error analysis showed that after augmenting the text, each score level had very little overlaps with the adjacent levels (Table 13). For example, the Elementary level which had the highest overlap with the Pre-Intermediate in the pre-augmentation stage contained little overlap (n=16) with the adjacent level after adding the augmented text. Also, the Upper-Intermediate and the Advanced levels that had the highest number of overlaps before text augmentation, have relatively little overlap when adding more data to the score level. However, as the confusion matrix before augmentation (Table 9) showed, when the levels are more distant, the amount of classification overlap is less. For example, just one of the advanced level essays was classified as elementary using the augmented data.

Table 13.

Confusion Matrix by Level After Text Augmentation

Level	Elementary	Pre-Intermediate	Intermediate	Upper-Intermediate	Advanced
Elementary	**610**	16	11	3	4
Pre-Intermediate	12	**541**	13	4	2
Intermediate	10	34	**366**	29	17
Upper-Intermediate	3	3	10	**265**	6
Advanced	1	14	0	35	**237**

Chapter Summary

In this chapter the results from my evaluation of the Persian AES system were presented. I started by providing an overview of the processes that I conducted to obtain and implement the data for the current study including ethics requirements and the data pre-processing techniques. Next, I described the implementation of the AES architectures including the baseline Word2vec model and the BERT AES model. The statistical performance of both AES models was presented and compared with one another. The BERT AES model produced high reliability with human raters (QWK=0.84 vs Word Embedding Baseline QWK=0.75). The performance of the model was the highest at the Elementary level (QWK=0.80 vs Baseline QWK=0.68) and the lowest at the Advanced level (QWK= 0.56 vs Baseline QWK=0.54). The κ reliability was also high (κ= 0.93 vs Baseline κ=0.82) except for the Advanced level (κ=0.71 vs Baseline κ=0.67). The results from the accuracy measures showed that 73% of the total number of the essays were correctly scored by the BERT AES model compared to 71% for the Baseline AES model. Given the classification consistency at each score level, a technique—text augmentation—was implemented that could be used to improve the BERT AES model classification results. Augmentation was implemented and the classification results for the BERT AES model were

improved. The original BERT AES model (QWK= 84%, κ=88%) was improved after including the augmented texts (QWK=96%, κ=96%). Likewise, the performance of the model at each score level was improved after text augmentation. The model performance was improved the most at the Advanced level (QWK +0.35) and the least at the Elementary level (QWK +0.16). The results from the accuracy measures after text augmentation shows that 91% (vs. 73% for the original BERT AES model) of the total number of the essays were correctly scored by the system.

CHAPTER 5

DISCUSSION

This chapter begins with an overview and summary of some of main concepts that guided this study. The main findings of the study are then summarized. Next, I discuss the limitations of the study and the directions for future research. The chapter concludes with a presentation of implications of this research for practice.

Purpose of the Study

Persian as a Unique Language for Automated Essay Scoring

The purpose of this study was to develop, describe, and evaluate the first AES system for scoring essays using the Persian language. Persian is an Indo-European language which is spoken by more than 110 million people. The unique writing and linguistic system of the Persian language necessitate the development of a specific and unique AES system for three reasons. First, the alphabetic system of Persian is different from English meaning that the majority of the NLP libraries currently in use are not applicable to Persian. Moreover, Persian's standard orthography makes use of a combination of spaces and semi-spaces (zero-width non-joiners). Another noticeable difference in Persian orthography is that, unlike English, Persian is written from right to left. Second, Persian has a complex morphological system that contains a heavily suffixing affixational morphology with no expression of grammatical gender (Amtrup et al., 2000). Hence, suffixes in Persian pose important challenges in automated text analysis because one suffix can have many different meanings. In addition, adjectives in Persian have a limited inflection space. As a result, they may be simple, comparative, or superlative. In comparative and superlative forms, a suffix attaches to the adjective: '+تر (pronounced/tar/=er [in English]) for comparative and '+ترین (pronounced/tari:n/=est [in English]) for superlative adjectives. This

application differs from English which uses both suffixes (‘+er/+est’) and multi-word construction with ‘more/most’, in addition to some irregular cases such as ‘good’, ‘better’, and ‘best’. Third, verbs in Persian may be inflected in different combinations for tense, mood, aspect, voice, and person. For example, the past tense stem is used with another auxiliary verb to create the future form. When an auxiliary verb is used, prefixes attach to the auxiliary verb instead of the root. The negative marker ‘+ن ‘ (pronounced /n/ [means ‘not’]) and the object pronouns are attached to the verbs producing more than 100 verb conjugated forms (Rasooli et al., 2013). In short, the AES system I created must be robust and generalizable in order to address these three unique writing and linguistic characteristics of the Persian language which differ markedly from the English language.

To date, the majority of the published AES studies have focused on essays written in English (Ramesh & Sanampudi, 2021). Studies on multilingual (i.e., languages other than English) AES, by comparison, are rare. Studies on Persian AES are non-existent. Hence, Persian is a unique case to study using language technology. Although Persian is a widely spoken language, automated text analysis in this language remains limited because Persian is a low-resource language in terms of the data that are available and the computational linguistic tools that have been used to study this language.

Transformers and the Importance of BERT

The recent development of transformers for language representation serves as a novel but promising method that can be used to solve the text analysis problem in Persian. Transformers are deep learning models containing a self-attention mechanism that learns the context of input text data thereby providing different weights to different parts of a text based on understanding the relationship between the sequence of inputs with the entire text (Vaswani et al., 2017). The

use of an attention mechanism for contextual understanding permitted the development of pre-trained language models such as BERT (BERT stands for **B**idirectional **E**ncoder **R**epresentation for **T**ransformers) (Devlin et al., 2018). BERT can be used to analyze text in multiple languages. I used multilingual BERT (see https://huggingface.co/bert-base-multilingual-cased) for my research as a state-of-the-art method to develop the first AES system for use in Persian.

BERT is a transformer-based encoder model for language representation that uses a multi-head attention mechanism and a bidirectional approach to learn the contextual relations between words and sentences in a text for accurate representation of the entire text. For text classification tasks like essay scoring, BERT has achieved the current state-of-art accuracy performance level among the currently available transformer models in English (Uto, 2021). The exceptional performance was achieved for two reasons.

First, BERT uses a bidirectional approach in learning the context. Unlike other language models that read the text input sequentially, the encoder in BERT reads the entire sequence of the input at once. This characteristic allows the model to learn the context of the input on the surrounding input which results in a more accurate performance of the system (Devlin et al., 2018). BERT uses two different training strategies called mask language modeling (MLM) and next sentence prediction (NSP) to learn the context of the input. MLM is predicting a missing word in a sentence. In MLM, the model randomly masks 15% of the words in a sentence and then runs the entire masked sentences through the model in order to predict the masked words. BERT also uses a second unsupervised task called next sentence prediction (NSP). NSP is predicting if one sentence naturally follows another. In NSP, the model learns to understand longer-term dependencies across sentences. Using the NSP strategy allows the model to predict each two-sentence sequence that follow one another in a text. BERT learns this knowledge by

receiving masked sentence embeddings that are concatenated in pairs as inputs during pre-training. Half of the embeddings are random and the other half are actual sentence pairs from the pool of training data.

Second, BERT is a self-supervised learning model that was developed and pre-trained by Google on the raw public available data, such as millions of sentences from the internet, without human intervention. Being pre-trained on a very large number of texts by Google, BERT contains 110 million parameters that are ready to be fine-tuned on a new training dataset. As a result, BERT contains rich structural information about language that enables it to achieve a high accuracy on task-specific classification even with limited amounts of training data. This unique characteristic of BERT is beneficial for text classification tasks where there are not enough data to conducted AES analyses using the conventional methods.

BERT and Multilingual AES

The vast majority of the applications of BERT have used English language corpora. Very few studies have been conducted on languages other than English to develop AES systems using BERT. Only two studies were reported in the literature. The first study was conducted by Hirao et al. (2020). They developed an AES system using a labeled dataset consists of 558 Japanese essays. They used the dataset to train three AES systems including BERT, LSTM, and machine learning (ML) with hand-crafted features (e.g., random forest). BERT achieved the highest accurate measure (QWK= 0.62) in essay scoring compared to the LSTM (QWK= 0.24) and the ML models (QWK= 0.52). The second study was conducted by Li and Dia (2020). They developed an AES system using a labeled dataset consists of 300 Chinese essays. Li and Dia (2020) used a BERT network to obtain the sentence vectors for essays and then used a bidirectional LSTM (Bi-LSTM) network with two layers to extract the essay vectors. The

researchers could not access a real training set that was written by Chinese students and scored by human raters, so they simulated 300 essays in Chinese for training their model. The model in which the BERT was combined with the Bi-LSTM outperformed (QWK= 0.60) the other models including the Bi-LSTM (QWK= 0.55) and the ML with hand-crafted features (QWK= 0.51).

Taken together, these two studies provide evidence that BERT-based AES systems can yield improvements over the other advanced neural network models, such as LSTM, for scoring essays written in languages other than English. However, the reported reliability in these two studies is low compared to those studies conducted on English language essays.

Persian Essays, Model Architecture, and Performance Measures

One reason for the comparatively low reliability may be attributed to the size of the datasets used in these two multilingual AES studies. Given the large number of parameters required in BERT, a small dataset used for tuning can easily be over fit leading to results with low accuracy (Ezen-Can, 2020). To address this problem, I analyzed data consisted of 2,000 Persian essays. The essays were written by non-native Persian language learners from the Saddi Foundation. The learners were female and male students from a diverse range of nationalities located in the Middle East, North Africa, South Asia, East Asia, North America, and Europe. The essay prompts required students to describe their ideas about different real-life situations (see Appendix A and B). The assigned word limit for each prompt was between 100 to 180 words and, across all of the essays in the dataset, the average length was about 164 words.

A holistic rubric was implemented to score the essays that covered general writing criteria including the variety and appropriateness of the vocabulary used, the variation and correct use of the tense of the sentences, sentence complexity, mechanics of writing, and the relevance and logical coherence of the sentences to the given prompt. The essays were scored (labeled)

holistically by the course instructors (raters). The raters were trained to evaluate the overall quality of the essays based on a holistic rubric and assign a single score to each essay in a range of 1 (Elementary) to 5 (Advanced). Each essay was scored by two raters and the average of the scores was reported. The holistic rubric covered general writing criteria including the variety and appropriateness of the vocabulary used, the variation and correct use of the tense of the sentences, sentence complexity, mechanics of writing, and the relevance and logical coherence of the sentences to the given prompt.

Students at the Elementary level had little or no exposure to Persian and had been learning basic reading, writing, and speaking skills. They were able to use memorized words and phrases and to express factual information by manipulating grammatical structure. Students at the Intermediate or Upper-Intermediate levels had understanding of the language to successfully express a wide range of relationships, such as temporal, sequential, cause and effect with a wider range of vocabulary use. Students at the Advanced levels had sufficient mastery of the language to shape their writing skill to address different purposes and to clearly defend or justify a particular point of view.

To analyze the Persian essays, I created and implemented a unique AES model architecture. The transformer neural network model created for this research was provided in Figure 4. The model consisted of BERT, a recurrent neural layer, and a classifier. After pre-processing the data, the sequence of inputs was fed into the pre-trained Multilingual BERT system. Depending on each essay length, various layers were be used for feature extraction. BERT generates embeddings for the words in a predefined window over the text. The predefined window length in this study was 256. Then, a recurrent network was used to aggregate the multiple embeddings into a single embedding after the transformer window swipes over the long text. The Long Short-

Term Memory (LSTM) model was used as the recurrent layer to aggregate the BERT output at each time step. This model combines the data into a sequence of vectors having the same length relative to their temporal position and temporal dependency with respect to the features in the essays. A classification neural network was then used to classify texts based on their representations. The classification neural network is a single layer fully connected neural network which is used to map the feature space into the essay score space. Using the SoftMax activation in this layer converts the networks output into a probability distribution over each essay score. For each input text, the selected essay score is the score level with the highest probability in the neural network output. In the final step of the model architecture, the combination of the BERT AES model and the recurrent and fully connected layers were trained end-to-end on the essay scoring dataset. Fine-tuning the base pre-trained BERT AES model optimized the features produced by this model to be more suitable for essay scoring task. The feature mapping learned by the BERT AES model in this step can cluster the essays with similar scores into the feature space.

Model validation in AES typically depends on comparing the similarity between the model performance and the human raters. Human judges are considered the gold standard and function as the explicit criterion for evaluating the performance of the AES system. Performance measures were used to evaluate agreement between the AES system and the human raters. These measures include the Quadratic Weighted Kappa (QWK), the Kappa coefficient, and error analysis. QWK and Kappa are used almost universally as consistency measures in AES (Shermis, 2014). QWK is a weighted Kappa score is used to overcome the problem inherent to Kappa, which is that it does not account for the degree of disagree. In addition to score consistency, performance can also be measured with an error analysis to evaluate score

prediction accuracy. The Kappa coefficient provides a chance-corrected index and is computed based on the ratio of the proportion of times the agreement is observed to the maximum proportion of times that the agreement is made while correcting for chance agreement (Siegel & Castellen, 1988). Accuracy is the number of classifications a model correctly predicts divided by the total number of predictions made. To understand the performance of the model in each score level, the error analysis of the model is produced. The outcome is called the confusion matrix. The confusion matrix helps to understand the misclassification rate or error rate when more than two scale points are classified. The matrix contains a measure of precision and recall as well as the F-score index. Precision refers to the number of true positives divided by the total number of positive predictions. Recall refers to the number of true positives divided by sum of true positives and false negatives. F-score combines the precision and recall of a model and it is defined as the harmonic mean of the model's precision and recall outcome.

Contribution of the Study

The purpose of this study was to expand the literature on multilingual AES by developing the first scoring system for the Persian language using BERT. This study serves as a noteworthy improvement over the two multilingual AES studies currently available in the literature by training the system on a sufficient number of essays. I also described and implemented a unique model architecture that can be used by future researchers and practitioners to model multilingual AES data using BERT. This study is significant because it focuses on an interdisciplinary research problem and uses sophisticated new methods in computing science to solve an essay scoring problem in educational testing. BERT is a novel text representation model with unique features that not only has the potential to increases the performance of text analysis systems, but also provide researchers and practitioners with analytic solutions to solve the problem of data

sparsity in low-resource languages. This study can be considered as one of the very first to use BERT to score essays in a low-resource language.

Another important contribution of the study is the dataset that will be used to train the multilingual BERT AES model is rare and unique. Studying a low-resource language requires enough available training data in order to model the linguistic characteristics of the data. Unlike the two existing studies on multilingual AES using BERT, this study used a sufficient number of scored essays to provide a reliable demonstration of the performance of my unique AES system.

Discussion of the Main Findings

Four main findings in this study are highlighted and discussed.

Finding 1: Emergence of BERT and the Importance of Expanding Multilingual AES

The emergence of transformers is now revolutionizing the field of NLP by dramatically improving the performance of a number of NLP tasks, such as machine translation, question answering, and language modeling. Transformers are encoder-decoder-based neural networks used to solve sequence-to-sequence problems by finding a mapping function f from an input sequence (e.g., word or sentence) of n vectors $X_{1:n}$ to a sequence of m target vectors $Y_{1:m}$. The seminal manuscript on this topic was published in 2017 by Vaswani et al. The application of transformers in NLP tasks was further accelerated one year later with the advent and release of a specific transformer-based language model called the Bidirectional Encoder Representation for Transformers or BERT. BERT is a transformer-based encoder model for language representation that uses a multi-head attention mechanism and a bidirectional approach to learn the contextual relations between words and sentences in a text for accurate representation of the entire text. The seminal manuscript on this topic was published in 2018 by Devlin et al. Hence, the years 2017 and 2018 are significant in the evolution of transforms and transformer models, respectively.

Of the 40 citations in Chapter 2 of my dissertation, 32 (80%) were from 2017 or later corresponding with the Vaswani et al. (2017) manuscript and 30 (75%) were from 2018 or later corresponding with the Devlin et al. manuscript. Twenty-four (60%) of the manuscripts cited in my literature review were published in the last four years. These outcomes demonstrate that the introduction of transformers, generally, and the BERT AES model, specifically, can be described as catalysts for research on multilingual AES. My research on the application of transformers to multilingual AES is built on these developments and uses the most recent methods; hence, the present research serves as both a novel and a new way of thinking about multilingual AES.

I also noted that only two studies have been published in the literature on multilingual AES using BERT. The studies were conducted by Hirao et al. (2020) in Japanese and Li and Dia (2020) in Chinese. Both studies provide evidence that BERT-based multilingual AES systems can produce reasonably accurate classification rates. However, the reported reliability in these two studies was low compared to those studies conducted on English language essays. One reason for the comparatively low reliability may be attributed to the size of the datasets used in each of these two AES studies. Hence, another important contribution of my study is the dataset used to train the multilingual BERT AES model. Conducting research on a low-resource language requires enough available training data to model the linguistic characteristics. Unlike the Hirao et al. (2020) and Li and Dia (2020) studies, I used enough labeled data to provide a reliable demonstration of the performance of my multilingual AES system.

Finding 2: Model Architecture

For this research, I designed, implement, and evaluated a unique architecture that can be used by future researchers and practitioners to model multilingual AES data using BERT. The model was presented in Figure 4. The model consisted of BERT, a recurrent neural layer, and a

classifier. BERT generates embeddings for the words in a predefined window over the text. The predefined window length in this study was 256. This number is determined based on the distribution of essays length in the proposed dataset. A recurrent network was used to aggregate the multiple embeddings into a single embedding after the transformer window swipes over the long text. The Long Short-Term Memory (LSTM) model was used as the recurrent layer to aggregate the BERT output at each layer and to pool the layers into a sequence of vectors which had the same length relative to their temporal position and temporal dependency with respect to the features in the essays. As a result, the aggregated 768-dimensional embedding captured the general information of each embedding that was required for the specific task of scoring essays written in Persian. A classification neural network was then used to categorize texts based on their representations in the 768-dimensional space. The classification neural network is a single layer fully connected neural network model that which is used to map the feature space into the essay score space. In the last step, the combination of the BERT AES model and the recurrent layers are then trained end-to-end on the essay scoring dataset. Fine-tuning the base pre-trained BERT AES model optimizes the features produced by this model to be more suitable for essay scoring task.

I also included a word embedding baseline model as a reference against which the result of the proposed BERT-based model were be evaluated. Word embedding is a technique for representing a word as a real number vector while preserving its meanings, semantic relationships, and alternative meanings by using distances among the words. Word embedding has been critical to improving the performance of various natural language processing tasks, such as syntactic parsing and sentiment analysis. Word2vec (Mikolov et al., 2011) is a commonly used word embedding technique, and it is often considered the standard for pre-trained word

embedding in text analysis. Thus, this method provided an excellent point of comparison for understanding the transformer model architecture introduced in my study.

Transformer models are available for more than 100 different languages. This language diversity makes transformer-based methods like the one I presented in this study easy to apply to a large set of languages with minor modifications. As a result, the architecture and methods described in this study can be easily adapted and used to score essays written in over 100 different languages, thereby supporting the application and wide-spread use of multilingual AES.

Finding 3: Performance of Multilingual BERT

The Word2Vec model served as the baseline in my research. Word2Vec served as an important point of reference for evaluating the performance of BERT because it provides a standard of comparison for how a word embedding method should perform. Relative to the baseline model, the results for the BERT AES model were impressive. The model consisting of BERT and LSTM performed with high classification consistency (QWK=0.84 vs. Baseline QWK=0.75). The performance of the model was the highest for the Elementary level (QWK=0.80) and the lowest for the Advanced level (QWK= 0.56). The κ reliability was also very strong. The BERT AES performed with high reliability (κ= 0.93 vs. Baseline κ= 0.82) and reached a level described by Koch and Landis (1977) as "almost perfect agreement" (p. xx add citation for this quote here). The κ for each level also shows that the machine and human raters' score agreement was high (>0.80) for each level except for Advanced (κ=0.71). The result from the accuracy measures shows that about 73% of the total number of the essays were correctly scored by the BERT AES model. Of those essays that were considered as correctly classified by the AES system, more than 70% in each level except for Advanced were scored the same by the human raters (i.e., true positive). Among the essays that were incorrectly classified, more than

70% of them in each score level—except for Advanced—were considered as incorrect classifications (i.e., true negative). Error analysis showed that each level had some overlap with the adjacent levels, with the Upper-Intermediate and the Advanced levels having the highest number of overlaps. Taken together, these results demonstrate that the BERT AES model can be used with a high degree of precision to predict the essay scores produced by the raters in this study.

The performance of the BERT AES model was also higher than that of the existing multilingual AES studies using BERT (see Hirao et al., 2020; Li & Dai, 2020). This improvement can be attributed to the amount of data used in each study. Given that the models in Hirao et al. (2020) and Li and Dai (2020) studies were also based on LSTM and BERT, but with a significantly smaller amount of training data, I conclude that deep learning models including BERT need a large amount of data to perform consistently and reliably.

Taken together, these results demonstrate that the BERT AES model is very accurate at producing a substantial level of agreement between the AES system and the human raters for scoring essays written in Persian. With an adequate essay sample, the model can perform at a high level of consistency. Finding 4: Data Augmentation and Multilingual AES with Low-Resource Languages

One consistently occurring finding in this study was the differential performance across the essay score levels. The BERT AES model produced the most accurate results at the Elementary level and lowest at the Advanced level. In fact, the results at the Advanced level were consistently weak. For example, the Advanced level yielded a QWK of 0.56 (model average 0.84), a Kappa of 0.71 (model average 0.93), Precision score of 0.63 (model average 0.72), a Recall score of 0.62 (model average 0.73), and a F-Score of 0.62 (model average 0.72). In

addition, the results from the error analysis in the Confusion matrix reveal that essays were most commonly misclassified at the Upper-Intermediate level and its adjacent levels. One reason for the comparatively weak results at the higher performance levels may be attributed to the descriptors used with the holistic rubric. Perhaps the descriptors were too similar between levels. When there are small differences between rating levels, the machine usually needs more or different types of data to understand the granularity and nuances between levels that are created to be similar to one another. Another reason could be attributed to the number of essays at these upper levels. Perhaps the Upper-Intermediate and the Advanced levels did not contain enough data for the AES system to learn the distinction between content in those two score levels. To evaluate this interpretation and to attempt to correct this problem, text augmentation was used.

The conditional data augmentation method described by Kumar et al. (2021) was implemented in the current study to address the data sparsity problem in the Persian dataset. To implement this augmentation method, the results of the current study were expanded by using the pre-trained BERT to augment 20% of the original data at each score level. Thirty percent of the tokens were masked and 30% of the token chunks (window size 6) in each text were produced for the Elementary, Pre-Intermediate, Intermediate, and Upper-Intermediate levels. BERT was trained to replace the masked tokens and chunks with the most similar tokens and chunks in texts with the same score level. By replacing parts of tokens and sentences in a text, new data was created to augment the training and testing datasets.

Augmentation has a significant impact on the classification results. Overall, QWK for the original model (QWK= 84%) was improved after including the augmented texts (QWK=96%)—a noteworthy increase of +0.12. More importantly, the QWK performance of the model at each level was improved after data text augmentation. The model performance was improved the most

at the Advanced level (+0.35) and the least at the Elementary level (+0.16). Improvements were also noted for the Pre-Intermediate (+0.28), Intermediate (+0.22), and Upper-Intermediate (+0.34).

The κ reliability also improved markedly. The augmented BERT AES model performed with high reliability (κ= 0.96) compared to the original BERT AES model (κ= 0.88). The κ for each level also shows that the augmented BERT AES and human raters' score agreement was high (>0.80) for each level including for Upper-Intermediate (κ=0.84) and Advanced (κ=0.82). The result from the accuracy measures shows that 91% of the total number of the essays were correctly scored by the augmented BERT AES model. Of those essays that were considered as correctly classified by the augmented BERT AES system, 91%, on average, were scored the same by the human raters (i.e., true positive) including Upper-Advanced (89%) and Advanced (90%). Among the essays that were incorrectly classified, 90%, on average were considered as incorrect classifications (i.e., true negative) including Upper-Advanced (88%) and Advanced (89%).

Error analysis shows that each level had some overlap with the adjacent levels, but the Upper-Intermediate and the Advanced levels were comparatively small for the augmented BERT AES model relative to the original BERT AES model. These results demonstrate that augmentation can improve the accuracy of the BERT AES thereby providing a high degree of precision to predict the essay scores produced by the raters in this study. These results allow me to conclude that, in contexts similar to the one for this study, augmentation is an appropriate strategy for improving differential performance across the essay score levels, particularly for low-resource languages like Persian. The augmented BERT AES model resulted in noticeable more accurate results at the Upper-Intermediate and Advanced levels.

Limitations and Directions for Future Research

Use of One Language and One Dataset

The purpose of this study was to develop an AES system to score written-response tasks in a low-source language. Persian, although widely spoken, is considered a low-resource language with respect to the availability and use of data and computational linguistic tools (e.g., Habib, 2021; Khashabi et al., 2021). As a result, automated text analysis in Persian remains limited. Moreover, the unique writing and linguistic system of the Persian language necessitates the development of a specific and unique AES system. The current study used a very unique dataset. I analyzed the results from 2,000 Persian essays. The essays were written by non-native Persian language learners from the Saddi Foundation which is a Persian language education center in Iran. The learners were female and male students from a diverse range of nationalities located in the Middle East, North Africa, South Asia, East Asia, North America, and Europe.

The first important limitation of this study is the focus on a single low-resource language—Persian—and the use of one dataset. Hence, one direction for future research would be to evaluate other low-resource languages using the model architecture introduced in my study in order to determine the generalizability of the data and the model I evaluated.

Limitations of Scoring Rubric

The holistic rubric covered general writing criteria including the variety and appropriateness of the vocabulary used, the variation and correct use of the tense of the sentences (e.g., the use of active and passive sentences, subject/verb agreement), sentence complexity (e.g., the correct use of simple and complex sentences), mechanics of writing (punctuations and spelling), and the relevance and logical coherence of the sentences to the given prompt (i.e. the topic that students were asked to write about). The essay prompts required students to describe their ideas about

different real-life situations. For example, “describe your student life experience at the dormitory”, “how children should spend their free time”, and “describe your idea about eating at home or eating out” (see Appendix A and B). The assigned word limit for each prompt was between 100 to 180 words. Across all of the essays in the dataset, the average length was about 164 words. The students were assigned a score from 1-5 that corresponded to the proficiency levels described as Elementary, Pre-Intermediate, Intermediate, Upper-Intermediate, and Advanced. Students at the Elementary level had little or no exposure to Persian and had been learning basic reading, writing, and speaking skills. They were able to use memorized words and phrases and to express factual information by manipulating grammatical structure. Learners at the Intermediate or Upper-Intermediate levels had understanding of the language to successfully express a wide range of relationships, such as temporal, sequential, cause and effect with a wider range of vocabulary use. Learners at the Advanced levels had sufficient mastery of the language to shape their writing skill to address different purposes and to clearly defend or justify a particular point of view.

The second important limitation of this study is the rubric used in this study. The number of students at each level was not equal. The highest number of essays were written by students at the Elementary level of language proficiency and the lowest number of essays were written by the advanced students. The error analysis revealed that while each level had some overlap with the adjacent levels, the Upper-Intermediate and the Advanced levels had the highest number of overlaps comparing with their overlaps with other levels. As a result, the AES model accuracy performance was the lowest at the Upper-Intermediate and Advanced levels. I conclude that this result occurred because the descriptors used with the holistic rubric were too similar between levels. When there are small differences between rating levels, the machine requires more data in

addition to different types of data to understand the granularity and nuances between levels that are similar to one another. Hence, another direction for future research would be to obtain scored essays from Persian or another low-resource language that contains a more discriminating rubric that allows me to clearly differentiate the outcomes at each score level. I addressed this problem in the current study by augmenting the data. But augmentation would not be necessary if we could analyze data that contained a more discriminating rubric.

Limitation of the Raters

The essays were scored holistically by the course instructors. The raters were trained to evaluate the overall quality of the essays based on a holistic rubric and assign a single score to each essay in a range of 1 (Elementary) to 5 (Advanced) (see Table 1). To reduce the subjectivity of holistic scoring, each essay was scored by two raters and the average of the scores was reported by the Saddi Foundation. Therefore, the final score used in this study was based on the average of the two scores for each essay. Unfortunately, data on rater reliability was not available. As a result, the reliability of the rater's scores is unknown. In future research, more information could be collected on the reliability of the rater's scores. This information would help me understand the consistency of the scores that produced the final essay score. It may also help account for the inconsistency in the scores Upper-Intermediate and Advanced levels if, for example, rater reliability turned out to be lower at the high end of the score scale compared to the lower and middle sections of the score scale.

Other Methods of Data Augmentation

One important contribution of this study was to demonstrate the benefit of data augmentation. I reported that the BERT AES model produced differential performance across the essay score levels with the most accurate results at the Elementary level and the most inaccurate results at the

Advanced level. One reason for the comparatively weak results at the higher performance levels may be the descriptors used with the holistic rubric. Perhaps the descriptors were too similar between levels. Another reason could be attributed to the number of essays at these upper levels. Perhaps the Upper-Intermediate and the Advanced levels did not contain enough data for the AES system to learn the distinction between content in those two score levels. To address the possibility of my second interpretation, text augmentation was used. I implemented the conditional data augmentation method described by Kumar et al. (2021). I demonstrated that this augmentation method had a significant positive impact on the classification results. Overall, the original model was improved after including the augmented texts using all of the performance measures in my study. The results provide a comprehensive demonstration that the augmentation can improve the accuracy of the BERT AES thereby providing a high degree of precision to predict the essay scores produced by the raters in this study. I also concluded that augmentation is an excellent strategy for improving differential performance across the essay score levels, particularly for low-resource languages like Persian.

While the augmentation findings were positive, I only included the conditional data augmentation method described by Kumar et al. (2021). Hence, another limitation of my study was the focus on one augmentation method. Other methods are available, including but not limited to synonym replacement, random insertion, random swap, and random deletion, lexical-based replacement, and embedding replacement. For future research, I suggest that other data augmentation methods be evaluated using low-resource languages to determine the benefits and drawbacks of these approaches relative to the conditional data augmentation method presented in the current study.

Implications for Practice

Automated essay scoring (AES) is a method where a machine attempts to provide scoring decisions by emulating how the written responses were graded by human raters. This method is gaining popularity because of the rapidly expanding use of online learning courses where it can be applied used to efficiently and economically score students written tasks. In addition, the rapid development of neural network models and NLP techniques has increased the accuracy of AES methods to the point where the reliability of the AES system is comparable to the scores provided by human raters.

It is perhaps surprising to learn that, to date, the majority of the published AES studies have focused on essays written in English. Studies on multilingual AES, by comparison, are practically non-existent. Multilingual AES is a critically important research area because the language of assessment for many students throughout the world is not English. For instance, the Programme for International Student Assessment (PISA) is administered in over 90 different languages.

In the current study, I focused on Persian—an Indo-European language spoken by more than 110 million people around the world. The need to develop an AES system for Persian is an important problem to solve because Persian is the language of instruction and assessment at schools and higher education institutions in different countries. For example, the UNESCO Institute for Statistics reports that more than 60% of the population in Iran are students and the enrollment rate at the elementary school level is very high (UNESCO Institute for Statistics, 2015). Hence, students from Iran and other Persian speaking countries can benefit from the develops in AES. To address the challenges inherent to scoring a Persian written-response task, I obtained a large language-specific data set that was used to train a state-of-the-art AES system in

order to learn the varieties in the written tasks using multilingual BERT. My system was very accurate in scoring the essays and yielding results that very consistent with the scores produced by human raters.

But much work still remains. In a forthcoming special issue in the *International Journal of Artificial Intelligence in Education* titled "Educational NLP for a Multilingual World: Research, Applications, and Challenges", the guest editors Giora Alexandron, Beata Beigman Klebanov, Mamoru Komachi, and Torsten Zesch provided this sobering claim in the call for paper in this special issue:

> *While in English the past decade has seen substantial scientific progress in these directions, in languages other than English research is much more limited and lags far behind. If this trend continues, the digital learning technologies that will be available for learners in languages other than English will be inferior to those available for English-speaking learners, widening already existing global gaps between learners.*

Advances in AES such as those described in the present research can provide non-English-speaking students from around the world with access to the same educational opportunities as English-speaking students. One small component of these opportunities is access to high-quality educational tests. The purpose of my dissertation research was to demonstrate how AES could be implemented as a high-quality written-response educational test in Persian. I created and evaluated a model architecture that can be used in Persian. The model was very accurate at scoring students' essays written in Persian. As a result, the methods and models presented in my research can now be used by language and educational institutions in Persian-speaking countries to score students' essays in an efficient manner and at a reasonable cost. As I noted earlier, transformer models are also available in more than 100 non-English languages. Because of the methods and outcomes described in the current study, transformer-based methods can be applied to a broad range of non-English languages with only minor modifications. Hence, the

architecture and methods described in this study can be adapted and used to score essays written in over 100 different languages thereby allowing language and educational institutions in non-English speaking countries to implement multilingual AES.

REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., & Zheng, X. (2015). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467.*

Abonizio, H. Q., Paraiso, E. C., & Barbon, S. (2021). Toward text data augmentation for sentiment analysis. *IEEE Transactions on Artificial Intelligence, 3(5),* 657-668.

Amtrup, J. W., Rad, H. M., Megerdoomian, K., & Zajac, R. (2000). Persian-English machine translation: An overview of the Shiraz project. *Memoranda in Computer and Cognitive Science* MCCS-00-319, NMSU, CRL.

Arrieta, A. B., Díaz-Rodríguez, N., Ser, J., Del Bennetot, A., Tabik, S., Barbado, A., et al. (2020). Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inform. Fusion 58*, 82–115. doi: 10.1016/j.inffus.2019.12.012

Attali, Y. (2013). Validity and reliability of automated essay scoring. In M.D. Shermis & J.C. Burstein (Eds.), *Handbook of Automated Essay Evaluation: Current Application and New Directions* (pp. 181-198). New York: Psychology Press.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V. 2. *The Journal of Technology, Learning and Assessment*, *4*(3).

Bayer, M., Kaufhold, M. A., & Reuter, C. (2022). A survey on data augmentation for text classification. *ACM Computing Surveys*. https://doi.org/10.1145/3544558

Berrar, D. (2019). Cross-validation. In S. Ranganathan, M., Gribskov, K., Nakai, C., Schönbach (Eds.), *Encyclopedia of Bioinformatics and Computational Biology* (pp. 542–545). Academic Press, Oxford.

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*(3), 727-746.

Cader, A. (2020, July). The potential for the use of deep neural networks in e-learning student evaluation with new data augmentation method. In *International Conference on Artificial Intelligence in Education* (pp. 37-42). Springer, Cham.

Camus, L., & Filighera, A. (2020). Investigating transformers for automatic short answer grading. In I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, & E. Millán (Eds.), *Artificial Intelligence in Education* (pp. 43-48). Springer, Switzerland.

Cao Y, Jin H, Wan X, & Yu Z (2020). Domain-adaptive neural automated essay scoring. In J. Huang et al. (Eds), *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 1011–1020). Association for Computing Machinery, NY.

Chanaa, A., & Faddouli, E. (2020). Predicting learners need for recommendation using dynamic graph-based knowledge tracing. In I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, & E. Millán (Eds.), *Artificial Intelligence in Education* (pp. 49-53). Springer, Switzerland.

Chernyavskiy, A., Ilvovsky, D., & Nakov, P. (2021, September). Transformers:"The End of History" for Natural Language Processing?. *In Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 677-693). Springer, Cham.

Chung, G. K. W. K., & Baker, E. L. (2003). Issues in the reliability and validity of automated scoring of constructed responses. In M. D. Shermis, & J. C. Burstein (Eds.), *Automated Essay Scoring: A Cross Disciplinary Perspective* (pp. 23–40). Mahwah, NJ: Erlbaum.

Condor, A. (2020). Exploring automatic short answer grading as a tool to assist in human rating. In I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, & E. Millán (Eds.), *Artificial Intelligence in Education* (pp. 74-79). Springer, Switzerland.

Cui, X., Goel, V., & Kingsbury, B. (2015). Data augmentation for deep neural network acoustic modeling. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *23*(9), 1469-1477.

Davoudi, S., & Mirzaei, S. (2021, March). A Semantic-based feature extraction method using categorical clustering for Persian document classification. *In 2021 26th International Computer Conference, Computer Society of Iran (CSICC) (pp. 1-5). IEEE.*

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.* https://doi.org/10.48550/arXiv.1810.04805

Doewes, A., & Pechenizkiy, M. (2021). On the limitations of human computer agreement in automated essay scoring. In S. Hsiao, S. Sahebi, F. Bouchet, & J Vie (Eds), *Proceedings of the 14th International Conference on Educational Data Mining* (pp. 475-480). International Educational Data Mining Society, Paris.

Dong, F., Zhang, Y., & Yang, J. (2017). Attention-based recurrent convolutional neural network for automatic essay scoring. In R. Levy & L. Specia (Eds), *Proceedings of the 21st Conference on Computational Natural Language Learning* (pp. 153-162). Association for Computational Linguistics, Canada.

Ezen-Can, A. (2020). A comparison of LSTM and BERT for small corpus. *arXiv preprint arXiv:2009.05451.* https://doi.org/10.48550/arXiv.2009.05451

Feng, Z., Zhou, H., Zhu, Z., & Mao, K. (2022). Tailored text augmentation for sentiment analysis. *Expert Systems with Applications*, 117605.

Firoozi, T., Naeimabadi, A. N., Demmans Epp, C.,Bulut, O., Barbosa, D. (2022, April). *The effect of word vector representation and linguistic features on the accuracy of automated essay scoring systems using neural networks* [Paper presentation]. National Council on Measurement in Education (NCME) Conference 2022, San Diego, CA.

Firth, J. R. (1962). A synopsis of linguistic theory, 1930-1955. *Studies in Linguistic Analysis.*

Gierl, M. J., Latifi, F., Lai, H., Boulais, A-P, & De Champlain, A. (2014). Automated essay scoring and the future of assessment in medical education. *Medical Education, 48,* 950-962.

Habib, M. K. (2021). The challenges of Persian user-generated textual content: A machine learning-based approach. *arXiv preprint arXiv:2101.08087.*

Hale, S., & Campbell, S. (2002). The interaction between text difficulty and translation accuracy. *Babel, 48*(1), 14-33.

Haller, S., Aldea, A., Seifert, C., & Strisciuglio, N. (2022). Survey on automated short answer grading with deep learning: from word embeddings to transformers. *arXiv preprint arXiv:2204.03503.* https://doi.org/10.48550/arXiv.2204.03503

Hellman, S., Rosenstein, M., Gorman, A., Murray, W., Becker, L., Baikadi, A., & Foltz, P. W. (2019, June). Scaling up writing in the curriculum: Batch mode active learning for automated essay scoring. In *Proceedings of the Sixth (2019) ACM Conference on Learning at Scale* (pp. 1-10).

Hirao, R., Arai, M., Shimanaka, H., Katsumata, S., & Komachi, M. (2020). Automated essay scoring system for nonnative Japanese learners. In N. Calzolari et al. (Eds), *Proceedings of*

the 12th Language Resources and Evaluation Conference (pp. 1250-1257). European Language Resources Association, Paris.

Hayashi, T., Watanabe, S., Zhang, Y., Toda, T., Hori, T., Astudillo, R., & Takeda, K. (2018, December). Back-translation-style data augmentation for end-to-end ASR. *In 2018 IEEE Spoken Language Technology Workshop* (SLT) (pp. 426-433). IEEE.

Hochreiter, Y. Bengio, P. Frasconi, J. Schmidhuber (2001). Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. In S.C. Kremer, J.F. Kolen (Eds.), *A Field Guide to Dynamical Recurrent Neural Networks*, IEEE Press.

Hui, B., Liu, L., Chen, J., Zhou, X., & Nian, Y. (2020). Few-shot relation classification by context attention-based prototypical networks with BERT. *EURASIP Journal on Wireless Communications and Networking, 2020* (1), 1-17.

Jang, B., Kim, M., Harerimana, G., Kang, S. U., & Kim, J. W. (2020). Bi-LSTM model to increase accuracy in text classification: Combining Word2vec CNN and attention mechanism. *Applied Sciences, 10*(17), 5841.

Jiang, W., Zhang, K., Wang, N., & Yu, M. (2020). MeshCut data augmentation for deep learning in computer vision. *PLoS One*, *15*(12), e0243613.

Jong, Y. J., Kim, Y. J., & Ri, O. C. (2022). Improving performance of automated essay scoring by using back-translation essays and adjusted scores. *Mathematical Problems in Engineering.*

Jungiewicz, M., & Smywiński-Pohl, A. (2019). Towards textual data augmentation for neural networks: synonyms and maximum loss. *Computer Science*, *20*.

Ke, Z., & Ng, V. (2019). Automated essay scoring: A survey of the state of the art. In S. Kraus (Ed) *Proceedings of the Twenty-Eighth International Joint Conference on Artificial*

Intelligence (pp. 6300-6308). International Joint Conferences on Artificial Intelligence, Macao.

Khashabi, D., Cohan, A., Shakeri, S., Hosseini, P., Pezeshkpour, P., Alikhani, M., ... & Yaghoobzadeh, Y. (2021). Parsinlu: a suite of language understanding challenges for Persian. *Transactions of the Association for Computational Linguistics, 9*, 1163-1178.

Khayi, N. A. (2021). *Advancement auto-assessment of students' knowledge states from natural language input* [Doctoral dissertation]. The University of Memphis.

Kim, Y. (2014). Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882.*

Klebanov, B. B., & Madnani, N. (2022). Automated essay scoring. *Synthesis Lectures on Human Language Technologies, 14(5),* 1-314.

Ko, T., Peddinti, V., Povey, D., & Khudanpur, S. (2015). Audio augmentation for speech recognition. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Kobayashi, S. (2018). Contextual augmentation: Data augmentation by words with paradigmatic relations. *arXiv preprint arXiv:1805.06201.*

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84-90.

Kumar, V., & Boulanger, D. (2020, October). Explainable automated essay scoring: Deep learning really has pedagogical value. *Frontiers in Education, 5*.

Kumar, V., Choudhary, A., & Cho, E. (2021). Data augmentation using pre-trained transformer models. *arXiv preprint arXiv:2003.02245*.

Lagouvardos, S., Dolby, J., Grech, N., Antoniadis, A., & Smaragdakis, Y. (2020). Static analysis of shape in TensorFlow programs. In R. Hirschfeld & T. Pape (Eds), *34th European Conference on Object-Oriented Programming*. Schloss Dagstuhl – Leibniz-Zentrum für Informatik GmbH, Dagstuhl Publishing, Saarbrücken/Wadern, Germany.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174.

Latifi, S. (2016). *Development and validation of an automated essay scoring framework by integrating deep features of English language* [Unpublished doctoral dissertation]. University of Alberta. Education & Research Archive (ERA). https://doi.org/10.7939/R37S7J134

Levy, O. and Goldberg, Y.(2014, June). Dependency-based word embeddings. *In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (Volume 2: Short Papers) (pp. 302-308).

Li, Y., Pan, Q., Wang, S., Yang, T., & Cambria, E. (2018). A generative model for category text generation. *Information Sciences*, 450, 301-315.

Li, H., & Dia, T (2020). Explore deep learning for Chinese essay automated scoring. *Journal of Physics: Conference Series, 1631* (1).

Liu, Q., Kusner, M. J., & Blunsom, P. (2020). A survey on contextual embeddings. *arXiv preprint arXiv:2003.07278.* https://doi.org/10.48550/arXiv.2003.07278

Lun, J., Zhu, J., Tang, Y., & Yang, M. (2020, April). Multiple data augmentation strategies for improving performance on automatic short answer scoring. *In Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, No. 9, pp. 13389-13396).

Marivate, V., & Sefara, T. (2020, August). Improving short text classification through global augmentation methods. *In International Cross-Domain Conference for Machine Learning and Knowledge Extraction* (pp. 385-399). Springer, Cham.

Masica, C. P. (1993). *The Indo-Aryan Languages*. Cambridge University Press.

Mayfield, E., & Black, A. W. (2020). Should you fine-tune BERT for automated essay scoring?. In J. Burstein et al. (Eds), *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 151-162). Association for Computational Linguistics, USA.

McNamara, D. S., Crossley, S. A., Roscoe, R. D., Allen, L. K., & Dai, J. (2015). A hierarchical classification approach to automated essay scoring. *Assessing Writing*, *23*, 35-59.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated Evaluation of Text and Discourse with Coh-Metrix*. Cambridge University Press.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781.* https://doi.org/10.48550/arXiv.1301.3781

Mikolov, T., Karafiát, M., Burget, L., Černocký, J., & Khudanpur, S. (2010). Recurrent neural network based language model. In *Eleventh Annual Conference of the International Speech Communication Association*.

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM, 38*(11), 39-41.

OECD (2021). *PISA for Development Assessment and Analytical Framework Reading, Mathematics and Science.* OECD publishing. https://www.oecd.org/pisa/sitedocument/PISA-2021

Oliveira, E., Alves, J., Brito, J., & Pirovani, J. (2019). The influence of NER on the essay grading. In R. Caldas, Y. Hu, L. Neto, & B. Markert (Eds), *International Conference on Intelligent Systems Design and Applications* (pp. 162-171). Springer, Cham.

Page, E. B. (1994). Computer grading of student prose, using modern concepts and software. *The Journal of Experimental Education*, *62*(2), 127-142.

Page, E. B. (1966). The imminence of grading essays by computer. *The Phi Delta Kappa, 47*(5), 238-243.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (pp. 1532– 154). Association for Computational Linguistics, Stroudsburg, PA.

Pires, T., Schlinger, E., & Garrette, D. (2019). How multilingual is multilingual BERT?. *arXiv preprint arXiv:1906.01502.* https://doi.org/10.48550/arXiv.1906.01502

Qin, Y., Du, J., Wang, X., & Lu, H. (2019). Recurrent layer aggregation using LSTM. In *International Joint Conference on Neural Networks* (pp. 1-8). doi: 10.1109/IJCNN.2019.8852077.

Ramesh, D., & Sanampudi, S. K. (2021). An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review, 55* (3), 2495–2527.

Rasooli, M. S., El Kholy, A., & Habash, N. (2013, October). Orthographic and morphological processing for Persian-to-English statistical machine translation. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing* (pp. 1047-1051).

Rodriguez, P. U., Jafari, A., & Ormerod, C. M. (2019). Language models and automated essay scoring. *arXiv preprint arXiv:1909.09482*. https://doi.org/10.48550/arXiv.1909.09482

Roshanfekr, B., Khadivi, S., & Rahmati, M. (2017). Sentiment analysis using deep learning on Persian texts. In *Iranian Conference on Electrical Engineering* (pp. 1503-1508). IEEE, DOI: 10.1109/ICEE40693.2017

Rizos, G., Hemker, K., & Schuller, B. (2019, November). Augment to prevent: short-text data augmentation in deep learning for hate-speech classification. *In Proceedings of the 28th ACM international conference on information and knowledge management* (pp. 991-1000).

Scarlatos, A., Brinton, C., & Lan, A. (2022). Process-Bert: A framework for representation learning on educational process data. *arXiv preprint arXiv:2204.13607.* https://doi.org/10.48550/arXiv.2204.13607

Sennrich, R., Haddow, B., & Birch, A. (2015). Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Shehab, A., Elhoseny, M., & Hassanien, A. E. (2016). A hybrid scheme for automated essay grading based on LVQ and NLP techniques. In *12th International Computer Engineering Conference* (pp. 65-70). IEEE, Cairo, Egypt.

Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing*, 20(4), 53-76.

Shin, J., & Gierl, M. J. (2020). More efficient processes for creating automated essay scoring frameworks: A demonstration of two algorithms. *Language Testing, 38*(2), 247–272.

Siegel, S. C., & Castellan, J. (1988). *Nonparametric Statistics for the Behavioral Sciences.* New York, McGraw-Hill.

Simard, P. Y., LeCun, Y. A., Denker, J. S., & Victorri, B. (1998). Transformation invariance in pattern recognition—tangent distance and tangent propagation. In *Neural networks: Tricks of the trade* (pp. 239-274). Springer, Berlin, Heidelberg.

Simons, G. F., & Fennig, C. D. (2017). *Ethnologue: Languages of Asia.* Sil International: Dallas. TX.

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to fine-tune BERT for text classification?. In *China National Conference on Chinese Computational Linguistics* (pp. 194-206). Springer, New York.

Sung, C., Dhamecha, T., Saha, S., Ma, T., Reddy, V., & Arora, R. (2019). Pre-training BERT on domain resources for short answer grading. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing* (pp. 6071-6075). Association for Computational Linguistics, Hong Kong, China.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems, 27*.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1-9).

Taghipour, K., & Ng, H. T. (2016). A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 1882-1891). Association for Computational Linguistics, Austin, Texas.

Tseng, H. C., Chen, H. C., Chang, K. E., Sung, Y. T., & Chen, B. (2019). An innovative BERT-based readability model. In *International Conference on Innovative Technologies and Learning* (pp. 301-308). Springer, Cham.

UNESCO Institute for Statistics. (2015). *Adult and Youth Literacy.* http://www.uis.unesco.org/literacy/Documents/fs32-2015-literacy.pdf.

Uto, M. (2021). A review of deep-neural automated essay scoring models. *Behaviormetrika,48(*2), 459-484.

Urquhart, V.A., McIver, M.C. (2005). *Teaching Writing in the Content Areas.* Association for Supervision and Curriculum Development, Alexandria, VA.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30.* https://doi.org/10.48550/arXiv.1706.03762

Visa, S., Ramsay, B., Ralescu, A. L., & Van Der Knaap, E. (2011). Confusion matrix-based feature selection. In S. Vida (Ed), *22nd Midwest Artificial Intelligence and Cognitive Science Conference (pp.* 120-127). MAICS, Ohio, USA.

Wang, X., Sheng, Y., Deng, H., & Zhao, Z. (2019). CharCNN-SVM for Chinese text datasets sentiment classification with data augmentation. *International Journal of Innovative Computing, Information and Control, 15*(1), 227-246.

Wang, W. Y., & Yang, D. (2015, September). That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using# petpeeve tweets. *In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 2557-2563)

Wang, J., Yu, L. C., Lai, K. R., & Zhang, X. (2016, August). Dimensional sentiment analysis using a regional CNN-LSTM model. *In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short papers) (pp. 225-230).*

Wangkriangkri, P., Viboonlarp, C., Rutherford, A. T., & Chuangsuwanich, E. (2020). A comparative study of pretrained language models for automated essay scoring with

adversarial inputs. In *2020 IEEE Region 10 International Conference TENCON* (pp. 875-880). IEEE, Osaka, Japan.

Wei, J., & Zou, K. (2019). Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice, 31*(1), 2-13.

Wind, S. A., Wolfe, E. W., Engelhard Jr, G., Foltz, P., & Rosenstein, M. (2018). The influence of rater effects in training sets on the psychometric quality of automated scoring for writing assessments. *International Journal of Testing, 18*(1), 27-49.

World Economic Forum (2022, January). *Center for the New Economy and Society: These 3 charts show the global growth in online learning.*

Wu, X., Lv, S., Zang, L., Han, J., & Hu, S. (2019, June). Conditional BERT contextual augmentation. *In International Conference on Computational Science* (pp. 84-95). Springer, Cham.

Xue, J., Tang, X., & Zheng, L. (2021). A hierarchical BERT-based transfer learning approach for multi-dimensional essay scoring. *IEEE Access, 9*, 125403-125415.

Yu, A. W., Dohan, D., Luong, M. T., Zhao, R., Chen, K., Norouzi, M., & Le, Q. V. (2018). Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541*.

Zhang, M., Williamson, D. M., Breyer, F. J., & Trapani, C. (2012). Comparison of e-rater® automated essay scoring model calibration methods based on distributional targets. *International Journal of Testing, 12*(4), 345-364.

Appendix A

Essay Prompts in Persian

توضیحات:

- از 3 موضوع زیر را انتخاب کنید و دربارهٔ آن یک متن بنویسید. متن شما باید حدّاقل 180 واژه داشته‌باشد.
- برای نوشتن متن، 60 دقیقه زمان دارید.
- متن شما باید 3 تا 4 پاراگراف داشته‌باشد. در پاراگراف اول، مقدمه و کلیات را بیان کنید؛ در پاراگراف(های) میانی، ایده‌های خود را توضیح دهید و در پاراگراف آخر، نتیجه‌گیری کنید.

متون:

1) بعضی از افراد معتقدند باید کودکان اوقات فراغت خود را در کلاس‌های آموزشی بگذرانند، اما بعضی دیگر معتقدند آنها باید در اوقات فراغت خود آزاد باشند. این دو دیدگاه را با هم مقایسه کنید و نظر خودتان را بیان کنید.

2) بعضی از افراد معتقدند در انجام کارها و گرفتن تصمیمات، علم بیشتر از تجربه به انسان‌ها کمک می‌کند، اما بعضی دیگر تجربه را بهتر از علم می‌دانند. این دو دیدگاه را با هم مقایسه کنید و نظر خودتان را بیان کنید.

3) بعضی افراد اعتقاد دارند درآمد سیتاره‌های سینما و ورزش بیش‌ازحد بالاست. با این دیدگاه مخالف هستید یا موافق؟ :دیدگاه خود را توضیح دهید و دلایل خود را بیان کنید.

توضیحات:

- یکی از دو موضوع زیر را انتخاب کنید و دربارهٔ آن یک متن بین 10 تا 15 خط (100 تا 125 واژه) بنویسید
- برای نوشتن متن، 60 دقیقه زمان دارید.

متون:

1) آیا از اخبار کشور خود یا کشورهای دیگر جهان مطلع هستید؟ چگونه از اخبار مطلع می‌شوید؟ اینترنت، روزنامه، تلویزیون یا رادیو؟ خلاصهٔ یک خبر جالب را که به تازگی شنیده یا خوانده‌اید، بنویسید.

2) همان طور که می‌دانیم مردان و زنان هر یک توانایی‌ها و ضعف‌هایی دارند. آیا به نظر شما می‌توان افراد را به خاطر جنسیتشان از بعضی از شغل‌ها یا تحصیل در بعضی رشته‌های تحصیلی محروم کرد؟

توضیحات:

- یکی از موضوعات زیر را انتخاب کنید و دربارهٔ آن بنویسید. متن شما باید حدّاقل ۱۲۰ واژه داشته باشد.
- برای نوشتن متن، وقت دارید ۴۰ دقیقه زمان دارید.

متون:

1) دوستان می‌توانند در زندگی انسان تأثیر زیادی داشته باشند. به نظر شما یک دوست خوب باید چه ویژگی هایی داشته باشد؟ یکی از دوستان خود را توصیف کنید و ویژگی‌های اخلاقی او را بنویسید.

2) بعضی از مردم دوست دارند در خانه غذا بخورند چون به نظر آنها غذاهای رستوران آن‌ها را چاق و مریض می‌کند، ولی عده‌ای دیگر غذاهای رستوران را دوست دارند. چون خوشمزه‌تر است؛ به نظر شما کدام بهتر است؟ توضیح دهید.

3) دانشجویانی که در یک شهر یا کشور دیگر درس می‌خوانند، معمولاً در خوابگاه زندگی می‌کنند. زندگی در خوابگاه چگونه است و چه مشکلاتی دارد؟ زندگی در خوابگاه را با زندگی در خانه مقایسه کنید.

.

4) بسیاری از کارشناسان معتقدند که رسانه‌های گروهی (تلویزیون، اینترنت، رادیو و...) توانسته‌اند فکر و اندیشهٔ مردم را کنترل کنند؛ آیا شما با این موضوع موافق هستید؟ دلایل خود را بیان کنید. رسانه‌ها چه تأثیری بر مردم می‌گذارند؟

توضیحات:

- از 3 موضوع زیر را انتخاب کنید و دربارهٔ آن یک متن بنویسید. متن شما باید حدّاقل 180 واژه داشته‌باشد.
- برای نوشتن متن، 60 دقیقه زمان دارید.
- متن شما باید 3 تا 4 پاراگراف داشته‌باشد. در پاراگراف اول، مقدمه و کلیات را بیان کنید؛ در پاراگراف(های) میانی، ایده‌های خود را توضیح دهید و در پاراگراف آخر، نتیجه گیری کنید.

متون:

1) شما در دوران کودکی چه عادت‌هایی داشتید؟ الان چه عادت‌هایی دارید؟ این دو را با هم مقایسه کنید.

2) بعضی از افراد ترجیح می‌دهند کتاب‌های الکترونیکی بخوانند، اما بعضی دیگر هنوز خواندن کتاب‌های کاغذی را ترجیح می‌دهند. این دو دیدگاه را با هم مقایسه کنید و نظر خودتان را بیان کنید.

3) بعضی از مردم ترجیح می‌دهند برای فعالیت‌های روزمرهٔ خود برنامه‌ریزی کنند. اما برخی دیگر هیچ برنامهٔ خاصی برای کارهای روزمرهٔ خود ندارند. شما کدام‌یک را ترجیح می‌دهید؟ دلایل و مثال‌های خود را بنویسید و توضیح دهید

توضیحات:

- یکی از 3 موضوع زیر را انتخاب کنید و دربارهٔ آن یک متن بنویسید. متن شما باید حدّاقل 120 واژه داشته‌باشد
- برای نوشتن متن، 40 دقیقه زمان دارید.

متون:

1) در بعضی از کشورها بیشتر افراد در حال چاق شدن هستند و سلامتی آنها در خطر است. به نظر شما چه دلایلی باعث به وجود آمدن این مشکل شده‌است؟ برای حل این مشکل چه کارهایی باید انجام داد؟

2) بعضی از افراد معتقد هستند کسانی که برای تفریح و به مدّت کوتاه به کشورهای دیگر می‌روند، باید به فرهنگ آن کشورها احترام بگذارند و قوانین آن کشورها را رعایت کنند. بعضی یگر با این نظر مخالف هستند و معتقدند که کشور میزبان باید به تفاوت‌های بین فرهنگ‌ها و کشورها احترام بگذارد. این دو دیدگاه را با هم مقایسه کنید و نظر خودتان را بیان کنید.

3) بعضی از افراد معتقد هستند پدر و مادرها باید به کودکان یاد بدهند چگونه شهروندان خوبی برای جامعهٔ خود باشند. اما بعضی دیگر معتقدند مدرسه ها باید این کار را انجام بدهند. این دو دیدگاه را با هم مقایسه کنید و نظر خودتان را بیان کنید.

Appendix B

Essay Prompts Translated into English[1]

Instruction:

- Choose one of the following 3 topics and write a text about it. Your text should have at least 180 words.
- The allotted time for this task is 60 minutes.
- Your text should contain 3 to 4 paragraphs. In the first paragraph, state the introduction and generalities, in the middle paragraph(s), explain your ideas, and in the last paragraph mention the concluding remarks.

Topics:

1) Some people believe that children should spend their free time in educational classes, but others believe that they should be free to do whatever they want in their free time. Compare these two views and express your opinion.
2) Some people believe that science helps people more than self-experience in doing things and making decisions, but others believe that self-experience is better than science. Compare these two views and express your opinion.
3) Some people believe that the income of movie starts is unreasonably high. Do you agree or disagree with this view? Explain your point of view and state your reasons.

[1] The essay prompts are originally in Persian. The translations in English are provided by the researcher.

Instructions:

- Choose one of the two topics below and write a text about it.
- The text length should be between 10 and 15 lines (100 to 125 words).
- The allotted time for this task is 60 minutes.

Topics:

1) Do you know the news of your country or other countries of the world? What is your preferred source of news? Internet, newspaper, TV or radio? Write a summary of an interesting news story you recently heard or read.
2) As we know, men and women each have strengths and weaknesses. Do you think people can be excluded from certain jobs or study in certain fields of study because of their gender? Explain about your idea.

Instructions:

- Choose one of the following topics and write about it.
- Your text should contain at least 120 words.
- The allotted time for this task is 40 minutes.

Topics:

1) Friends can have a great influence on each other's life. What qualities do you think a good friend should have? Describe the characteristics of one of your friends and write down his/her moral qualities.
2) Some people like to eat at home because they think eating out makes them fat and sick, but some people like eating out because it is more delicious; Which do you think is better? Explain your idea.
3) Students who study in another city or country usually live in a dormitory. How is life in a dormitory and what are the potential challenges? Compare dorm life with home life.
4) Many experts believe that mass media (television, internet, radio, etc.) have been able to control people's thoughts; Do you agree with this idea? State your reasons. How does the media affect people?

Instructions:

- Choose one of the following topics and write a text about it.
- Your text should contain at least 180 words.
- The allotted time for this task is 60 minutes.
- Your text should contain 3 to 4 paragraphs. In the first paragraph, state the introduction and generalities, in the middle paragraph(s), explain your ideas, and in the last paragraph mention the concluding remarks.

Topics:

1) What habits did you use to have as a child? What are your habits now? Compare the two.
2) Some people prefer to read e-books, but others still prefer to read paper books. Compare these two views and express your opinion.
3) Some people believe that parents should teach children how to be good citizens in the society. But some others believe that schools are more responsible to educate children about this. Compare these two views and express your opinion.

Appendix C

Research Ethics Approval

Notification of Approval

Date:	Wednesday, January 25, 2023
Study ID:	Pro00127582
Principal Investigator:	Tahereh Firoozi
Study Supervisor:	Mark Gierl
Study Title:	Using Automated Procedures to Score Written Essays in Persian: An Application of the Multilingual BERT System
Approval Expiry Date:	Wednesday, January 24, 2024

Thank you for submitting the above study to the Research Ethics Board 2. Your application has been reviewed and approved on behalf of the committee.

Approved Documents:

Recruitment Materials
Candidacy paper

Any proposed changes to the study must be submitted to the REB for approval prior to implementation. A renewal report must be submitted next year prior to the expiry of this approval if your study still requires ethics approval. If you do not renew on or before the renewal expiry date, you will have to re-submit an ethics application.

Approval by the REB does not constitute authorization to initiate the conduct of this research. The Principal Investigator is responsible for ensuring required approvals from other involved organizations (e.g., Alberta Health Services, Covenant Health, community organizations, school boards) are obtained, before the research begins.

Sincerely,

Stanley Varnhagen, PhD
Associate Chair, Research Ethics Board 2

Note: This correspondence includes an electronic signature (validation and approval via an online system).