Journal of Computing in Higher Education (2021) 33:727–746
https://doi.org/10.1007/s12528-021-09283-1

ORIGINAL PAPER

A novel automated essay scoring approach for reliable higher educational assessments

Majdi Beseiso[1] · **Omar A. Alzubi**[1] · **Hasan Rashaideh**[1]

Accepted: 25 May 2021 / Published online: 1 June 2021

Abstract

E-learning is gradually gaining prominence in higher education, with universities enlarging provision and more students getting enrolled. The effectiveness of automated essay scoring (AES) is thus holding a strong appeal to universities for managing an increasing learning interest and reducing costs associated with human raters. The growth in e-learning systems in the higher education system and the demand for consistent writing assessments has spurred research interest in improving the accuracy of AES systems. This paper presents a transformer-based neural network model for improved AES performance using Bi-LSTM and RoBERTa language model based on Kaggle's ASAP dataset. The proposed model uses Bi-LSTM model over pre-trained RoBERTa language model to address the coherency issue in essays that is ignored by traditional essay scoring methods, including traditional NLP pipelines, deep learning-based methods, a mixture of both. The comparison of the experimental results on essay scoring with human raters concludes that the proposed model outperforms the existing methods in essay scoring in terms of QWK score. The comparative analysis of results demonstrates the applicability of the proposed model in automated essay scoring at higher education level.

Keywords Automated essay scoring (AES) · Deep learning · Essay scoring · Long short-term memory · Neural network · Transfer learning

✉ Majdi Beseiso
bsaiso@bau.edu.jo

Omar A. Alzubi
o.zubi@bau.edu.jo

Hasan Rashaideh
rashaideh@bau.edu.jo

[1] Computer Science Department, Al-Balqa Applied University, Salt, Jordan

Introduction

E-Learning systems are becoming increasingly known in higher education because of the flexible learning options demanded by the students and the financial pressures on higher educational institutions, which see technology as a way of reducing their expenses (Mahlangu, 2018). Essay scoring is considered as one of essential and critical job in the field of education. The assessment of students' knowledge based on essay type answers is a major concern in e-Learning systems. Manual scoring of essay is very time consuming and subjective to the experts. Automated essay scoring system enables grading the essay with the help of computer programs. This leads to reducing the work of manual scoring as well as the subjectivity of the experts. The growing deployment of e-Learning systems in higher education has led to numerous studies conducted for assessing and improving the consistency of the AES systems for efficient evaluation of high-grade essays (Hazar et al., 2019). AES has been a hot topic of research in academic literature for the past several years now. Higher education is presently challenged to handle the increasing number of learners with the existing learning model that exhibits growing evidence of excessive effort and cost and declining effectiveness of the learning process (Machicao, 2019). Under these circumstances, AES has reported several advantages in the educational setting, such as reliable reporting of scores, instant diagnostic feedback, and time-saving for teachers (Hussein et al., 2019). However, there is no human interaction and appreciation when AES is used in essay scoring. The system cannot recognize the artistic concept, emotion, novelty, and creativity in the submitted essays (Zhang & Wallace, 2015). Furthermore, these systems need huge datasets for training and are vulnerable to new types of test-taking tactics and cheating as well (Gierl et al., 2014). Moreover, the integration of AES is an open challenge, particularly, for universities having a low level of automation incorporated in their systems. The higher educational institutions are required to adapt and develop individual and organizational competencies, and transparency regarding the elements of new evaluation systems (Machicao, 2019). But despite these limitations and challenges, AES has continued to attract the attention of educators, researchers, testing companies, schools, and universities. A lot of work has been done lately to assess the application of AES in higher education (Citawan et al., 2018; Latifi et al., 2016; Li et al., 2020; Ng et al., 2019; Perin & Lauterbach, 2018; Report Series, 2012; Tobback, 2018; Yin et al., 2016; Zhang & Wallace, 2015). AES is gaining attention in the higher education sector as a viable approach for reducing resource-intensity of human scoring and achieving more consistent results compared to human raters. However, there is still room for improvement for the wide adoption of AES in this domain, which has led to numerous studies over the past years.

Several automated essay scoring systems have been developed in the recent past, including Intelligent Essay Assessor (IEA) (Foltz et al., 1999), Project Essay Grade (PEG) (Page, 1967), and E-rater (Attali & Burstein, 2006). These systems have been widely adopted in the field of education for automated essay scoring (Alikaniotis et al., 2016; Bennett & Bejar, 1998; Kwong et al., 2019; Ramineni et al., 2012; Yu & Barker 2020; Zhang, 2019). But these systems suffer from

different limitations and hence not considered promising in the coming time. Accuracy of any automated essay scoring system relies on the understanding of the language by the computer programs such as grammar, syntax, semantics and spellings in addition to essay grading rules and regulations.

Several traditional methods applied for developing an automated essay scoring system involve the use of machine learning techniques for classification problems (Larkey, 1998; Rong, 2014), and regression problems (Attali & Burstein, 2006; Phandi et al., 2015). Such approaches rely on manually extracted features such as essay length (Bernstein et al., 2010; Cushing Weigle, 2010; Ginther et al., 2010; Hartley, 2004). The major limitation of traditional automated essay scoring systems developed based on manually extracted features is that they are time-consuming systems and accuracy mainly depends upon the quality of the extracted features.

Recent advancement in computational power and neural network-enabled a massive potential developing efficient and effective Natural Language Processing systems. A neural network can be trained to learn an essay representation in the form of a single dense vector. A neural network can learn this dense vector representation for one to one mapping of essay scores. Recently deep learning techniques have advantages over traditional methods of extracting automatic features. Taking these advantages into consideration, several researchers have focused on deep learning methods to develop accurate automated essay scoring systems using neural networks (Alikaniotis et al., 2016; Bahdanau et al., 2014; Dong et al., 2017; Su et al., 2016; Taghipour & Ng, 2016).

Automated essay scoring systems based upon deep learning methods such as RNN have shown better performance in comparison to traditional systems (Kwong et al., 2019). Some ensemble-based methods have also been reported with improved accuracy in automated essay scoring (Zechner et al., 2009). The main focus of prior work was on the use of a conventional neural network (Santos & Gatti, 2014; Yang et al., 2016; Zhang, 2013), long short term memory (Hochreiter & Schmidhuber, 1997), recurrent neural network, and special intrinsic features like coherence features (Yang et al., 2016). Most of the prior work in the field of automated essay scoring suffers from the limitation is the use of a limited Corpus of properly written English text. Limited corpus of properly written English text lacks in understanding the context around words. To address this issue, word embedding methods like Word2Vec (Le & Mikolov, 2014) and GloVe (Pennington et al., 2014) have been developed. However, these methods do not capture the context while giving out the embedded representation for a particular word. This may lead to a big problem in essay scoring systems because an essay can be written in various forms and contexts for a given particular prompt. Besides, lookup table based embedding methods does not capture grammatical correctness contextual information of an essay unlike the RoBERTa encoder, which is considered as an essential scoring determinant. The limitation of these methods like word2vec and Glove is that they give the same embedding for words used in different contexts, they give the same embedding for fruit apple and the company Apple. LSTMs and RNNs have vanishing gradient problems and in case of essays, we have much longer input and gradient vanishing problem arises.

In this work, we use state of the art language understanding models, namely, RoBERTa (Liu et al., 2019) and XLNet (Wirth & Fu¨rnkranz, 2014), released by technology giants like Google, Face-book, and OpenAI along with LSTM to develop an accurate automated essay scoring system. These models have Transformers (Uzun, 2018) as their base unit of architecture that applies the concept of self-attention. Since these models have been pre-trained on a large corpus of standardized English text in an unsupervised manner, they can generalize well to the dataset at hand and can achieve state-of-the-art results. However, such models suffer from the limitation of document length truncation in the domain of essay scoring. Based on this assumption, we formulated an automated essay scoring problem as a regression problem. We proposed a novel solution using Bi-LSTM on the top of pre-trained RoBERTa model. The primary motivation of using Bi-LSTM on the top of RoB-ERTa model is to preserve extra recurrent structures that eliminate the problem of document length (essay length) for RoBERTa model. Integration of Bi-LSTM and RoB-ERTa models enable an analysis of the whole essay without leaving any part to achieve better scoring results.

Rest of the paper is structured as follows. Section 2 summarizes state of the art in the field of automated essay scoring systems along with their pros and cons. Section 3 automated essay scoring system proposed in this paper. It describes the pre-trained RoBERTa language model, its fine-tuning based on ASAP dataset, the proposed Bi-LSTM on the top of RobERTa model for automated essay scoring and its working. Section 4 describes the procedure of the experiment, including experimental setup, evaluation metrics, results and discussion. Finally, Sect. 5 concludes the paper at the end.

Related work

Recently, many research efforts have been invested in developing automated essay scoring systems (Bernstein et al., 2010; Kwong et al., 2019; Larkey, 1998; Phandi et al., 2015). Several leading educational agencies like Kaplan (Liu et al., 2019), Pearson (Machicao, 2019), ETS (Mahlangu, 2018) have played a significant role in this domain. In recent years, automated essay scoring has also gained research interest pertaining to its deployment in the higher education system (Machicao, 2019). Numerous studies have addressed the applications and challenges of AES in tertiary education (Citawan et al., 2018; Li et al., 2020; Ng et al., 2019; Zhang & Wallace, 2015).

Traditional methods like support vector machine focused on handcrafted linguistic features based automated essay scoring (Dascalu et al., 2014; Shi & Demberg, 2019; Wang et al., 2018). Whereas, open-source engine such as EASE (Rudner & Liang, 2002) formulated essay scoring problem as a regression problem. These approaches attempted to solve the problem by finding the ground truth scores and the predicted scores with the help of a complex regression loss. Advancement of word embedding methods like Glove (Pennington et al., 2014) and word2vec (Reilly et al., 2016) enable its usage and importance in successfully capturing the syntactic and semantic relationship between words. The use of word embedding methods

led to the development of end-to-end approaches using deep neural network architectures like LSTMs and CNNs (Huang et al., 2015; Nadeem et al., 2019). Several approaches have been proposed based on LSTMs and CNNs and their combination to solve the essay scoring problem as a classification problem (Nadeem & Ostendorf, 2018; Shermis & Hamner, 2013; Tay et al., 2017). For example, Kwong et al. (2019) used a simple LSTM based approach on low dimensional word embedding methods while Nadeem et al. (2019) combined CNN based features and LSTM based features for essay scoring.

Some researchers have also used LSTM and CNN models solving essay scoring problem as linear or logistic regression problem (Cummins & Rei, 2018; Farag et al., 2018; Su et al., 2016; Vaughn & Justice, 2015). The avenues of document classification have also explored in Nadeem and Ostendorf (2018) and Yang et al. (2019). However, most of these networks fail to achieve good results because they lack the sense of coherency in their understanding of language. To alleviate this problem, BERT and its variants (Bansal & Passonneau, 2018; Cer et al., 2018; Devlin et al., 2018) have been used to generate sentence embedding and utilized it hierarchically. However, no significant improvement over traditional models has been achieved. It can be attributed to the fact that although BERT and its variants were trained in an unsupervised manner on a massive corpus of free-flowing text, the pre-training objective is not perfect enough to understand the fluency, reasoning and lots of other characteristics that should be present in a well-written essay (given a prompt). BERT based approaches also suffer from the issue of document length truncation as limitation of BERT models. Besides, most of the prior work in the field of automated essay scoring suffers from the limitation is the use of a limited Corpus of properly written English text. Limited corpus of properly writ- ten English text lacks in understanding the context around words. Wang et al. (Vaughn & Justice, 2015) developed an essay scoring method using reinforcement learning algorithms. However, the problem with all these approaches includes the fact that none of these models was trained on large corpora of properly written English text that captures context around the words. This can be a massive problem because not capture the context while giving out the embedding for a particular word may lead to the wrong translation and scoring of the word. This may lead to a big problem in essay scoring systems because essay writers/candidates can write the essay in various forms and various contexts when given a particular prompt. Furthermore, grammatical correctness of an essay also determines the score assigned to an essay, which is again not captured by lookup table-based word embedding models.

It has been observed that prior research work in the field of natural language processing/understanding has well established BERT (Devlin et al., 2018) as state of the art in natural language processing tasks like sequence classification, question answering, etc. Thus, we use BERT model as a pure encoder for the classification task in this work. The primary motivation here is that BERT encoders have been trained in a better language modelling task and have a much better understanding of modelling the task-specific language corpus, which is very important in this case.

Therefore, we propose a transformer-based architecture using RoBERTa and Bi-LSTM models. The proposed architecture addresses the issue of document length truncation of BERT based models. Unlike HAN based net-works, this work

significantly changes the pre-training task of BERT models by giving more weightage to the next sentence prediction task using LSTM model. The proposed work is analogous to the work proposed by Farag et al. (2018). However, Farag et al. (2018) focused on modelling native text coherence for capturing adversarial examples and not for essay marking.

Automated essay scoring system

This work proposes an automated essay scoring system by integrating Bi-LSTM model on the top of RoBERTa model for exploiting the capability of temporal nature of LSTM model for addressing the issue of document length truncation of RoBERTa model. Here, we use RoBERTa model as the language encoder capture the coherency of the essay. In this section, we describe RoBERTa model and proposed model architecture.

RoBERTa model

RoBERTa is a large unsupervised pre-trained language model trained on 160 GB of text data. Like in BERT, the base unit of architecture is the Transformer. It was released by Facebook as an attempt to improve the performance of BERT. Figure 1 depicts the bird's-eye view of BERT model (Dascalu et al., 2014). RoBERTa and BERT have the same black-box architecture (Beseiso & Alzahrani, 2020). The output of this black box is a generalized contextual embedded representation. The main difference lies in the pre-training process of BERT and RoBERTa in which the latter uses just Masked Language Modeling (MLM) loss instead of a combination of MLM + NSP loss used in case of BERT. The improved performance was achieved by training RoBERTa on a larger corpus of text data than BERT and also with larger mini-batch size. RoBERTa differs from BERT in pre-training objective as well as Next Sentence Prediction (NSP) objective (Shermis & Hamner, 2013). Roberta was pre-trained using longer sentences and more number of epochs/Pre-training a language model longer results in better understanding of the language. We utilize this idea when we fine tune the Roberta model on out essay corpus. It helps the model

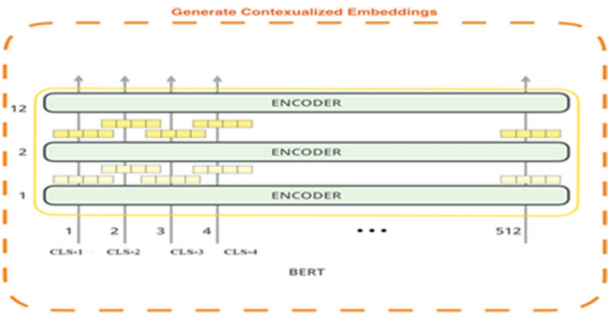

Fig. 1 A bird's-eye view of BERT (Dascalu et al., 2014)

get better understanding of essay domain. RoBERTa is only pre-trained on the Masked Language Model (MLM) objective (Lample & Conneau, 2019). Thus, it has a better understanding of language than the standard recurrent models like RNNs and LSTMs. The use of MLM objective enables RoBERTa model to capture the coherency in essays without caring about the sentence that comes next. This concept is particularly important in essays because essays are written based on a single prompt, and using an NSP loss along with MLM loss moves the model weights in such a way that the coherency changes at long periods in an essay (like para-graphs, etc.) get mostly penalized.

To alleviate this problem, RoBERTa model removed NSP loss and incorporated MLM objective. The MLM objective as the basis for training RoBERTa uniformly selects 15% of all tokens in the entire corpus as depicted in Fig. 2. This is in principal with Roberta working. All the BERT type models use this percentage. From the selected tokens, MLM objective requires replacing 80% of them with [MASK] token. 10% of the selected tokens are kept unchanged, and 10% are replaced with a random word. During pre-training time, the model then tries to predict the words at the [MASK] positions using the cross-entropy loss over the entire vocabulary.

The proposed model architecture

In this section, we introduce the overall architecture of the proposed model. Figure 3 presents the schematic representation of the entire architecture with the implementation point of view. The proposed model scores an essay with a given prompt.

The prompt refers to the topic of an essay. It acts as a part of the essay. With this assumption in mind, the proposed model tackles the problem of scoring essays with MLM loss by combining prompt and body of the essay. Therefore, it extracts a single chunk of text per essay. The proposed model involves fine-tuning of RoBERTa model on essay corpus. The fine-tuning is performed by minimizing MLM loss for nine epochs across the entire dataset. The MLM loss refers to the categorical cross-entropy loss over the entire vocabulary for the [MASK] tokens as presented in Fig. 4. Figure 4 shows the MLM loss as a combination of cross-entropy loss at multiple positions. The softmax operation is applied over the

Fig. 2 The Masked Language Modeling (MLM) objective as basis for training of RoBERTa model (Tay et al., 2017)

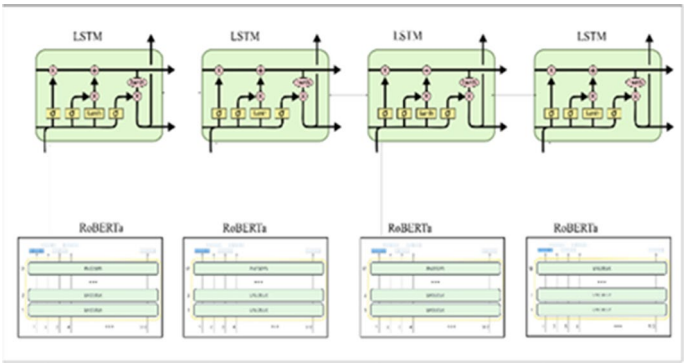

Fig. 3 The proposed model architecture

entire vocabulary and computed using Eq. 1. Here, the term (s)m represents the embedding representation, and f(s)m is the SoftMax operation over the entire vocabulary. This shows loss is minimized at every [MASK] token position. Thus, MLM loss is a combination of cross-entropy (CE) loss at multiple positions and computed as per Eq. 2.

$$f(s)_m = \frac{e^s_m}{P_{j_n e^s_n}} \tag{1}$$

$$CE = -\sum_{m}^{j} t_m \log(f(s_m)) \tag{2}$$

Fine tuning of Roberta model on essay corpus shifts the embedding space towards better understanding of essay writings. The MLM task on essay data helps the model learn the context in accordance with the essay sentences structure. Once the model is fine tuned on essay corpus, the attention heads in each Roberta layer tweak their weights so as to represent words in terms of essay context. This step plays a critical role in our method outperforming the other essay scoring methods. Fine tuning helps the model learn the language features in essay domain, how different words behave in essay writings.

The fine-tuned model on essay corpus is further used for demonstrating its validity using ASAP dataset. For this purpose, ASAP dataset is split into a ratio of 80:20 as training and test dataset. In this work, we propose to fine-tune the model initially and then followed by the splitting of the dataset. This process does not affect the performance of the model because the bias created by the test set on the weights of the model is negligible. After validating the model on test set, we fine tune the model on whole dataset as it is better practice in machine learning. Validation on test set has shown that validation dataset lies in the similar distribution as well so it is better to train on test set as well.

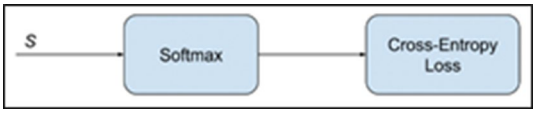

Fig. 4 A representation of MLM loss

To address the issue of document length restriction in RoBERTa model and its variants, we propose to divide the entire essay into lengths of 512 Byte-Pair tokens similar to the proposal in Taghipour and Ng (2016). Finally, the hidden state from the [CLS] or the first token of every chunk was exploited as the embedding representation for that chunk. The extracted embedded representation from each chuck is passed to Bi-LSTM model for maintaining the temporal sequence between them. As presented in Fig. 3, every RoBERTa block represents a chunk of document of length 512. Byte-Pair Tokens. The output of every RoBERTa block is a hidden state output from the [CLS] token or the first hidden state of every chunk. These representations are finally passed to the Bi-LSTMs to produce output as a final time step. Bi-LSTM is preferred over the other recurrent networks like RNNs and LSTMs because Bi_ LSTM is bi-directional whereas other models are unidirectional. Bi-LSTM looks at both the forward tokens and back tokens to understand the context of current tokens. Due to bidirectional nature, Bi-LSTM works better o longer sentences. In this model, the final embedding representation received from the last LSTM time step is used to compute the final score. This final score or predicted score is computed by inputting the final embedding representation through a bunch of fully connected layers and applying a sigmoid activation function at the end. Since sigmoid activation function is differentiable, so it takes care of normalization problems. The number produced as the output of this layer is considered as the predicted score of essay. The proposed model attempts to minimize mean squared error (MSE) loss over the actual scores and the predicted scores. The MSE Loss is computed as per Eq. 3.

$$\mathrm{MSE} = \frac{1}{\mathrm{T}} \sum_{\mathrm{j}=1}^{\mathrm{T}} (\mathrm{s_j} - \hat{\mathrm{s}}_\mathrm{j}) \tag{3}$$

where T is the total sample set size. The s_j gives the actual value or the ground truth, and the $\hat{s}_j$ shows the predicted value. The working of the proposed model is presented in Fig. 5 and summarized below.

Bi-LSTM are designed in such a way that they preserve the context in a directed manner. We have used [CLS] token to represent the mean of the whole sentence. It can be obtained using mean, max or any predefined pooling strategy. Mean pooling is the default pooling strategy and it serves our purpose well as we have to keep the coherency of essay intact. Mean vectors for different chunks when go through the LSTM network become more coherent and give us the final output when passed through the ANN. Doing so keeps the coherency and context intact in deep embedding space.

– Fine-tuning of RoBERTa language model on the essay corpus using ASAP dataset to obtain fine-tuned RoBERTa model, explicit for essay corpus.

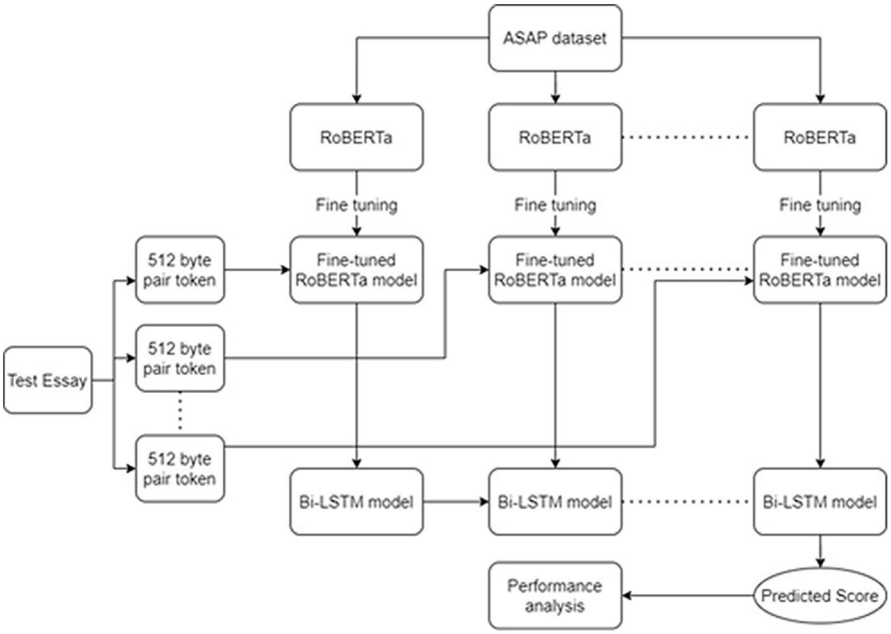

Fig. 5 Working of the proposed model

- Chunking test essay into document lengths of 512 byte-pair tokens.
- Feeding each chunk of 512 byte-pair tokens to fine-tuned RoBERTa models to get embedded representation corresponding to each chunk. The embedded representation is gathered from the [CLS] hidden state of the RoBERTa.
- The embedded representations of each chunk are input to Bi-LSTMs to maintain temporality. The hidden state output from the last LSTM time step was the reduced dimensionality using fully connected layers and the sigmoid activation function to produce the normalized numerical score for the test essay.
- Performance analysis module computes evaluation metrics for comprehensive comparison and performance analysis of the proposed model.

Experiments

In this section, we describe the procedure of the experiment, including experimental setup, evaluation metrics, results and discussion.

Experimental setup

For demonstrating the validity of the proposed model, we used ASAP dataset that is a Kaggle competition dataset (Uzun, 2018). ASAP dataset is sponsored by the William and Flora Hewlett Foundation (Hewlett Foundation) in 2012. This dataset

is extensively used as a benchmark dataset and enable to compare performance with state of the art in the field of essay scoring. Table 1 summarizes the prompts and genre of the ASAP dataset.

We trained our model for 180 epochs. The reason behind choosing 180 epochs is that the language model Roberta has already been fine tuned in essay dataset. Training the model longer than 180 epochs will result in over-fitting. All the other models has been trained for 180 epochs and in the results sections we will see our model converging much faster than the other models with a better performance in terms of mean squared error.

This dataset consists of essays written by seventh to 10th-grade students in ASCII format. Every essay has multiple gold annotations associated with it. Hu- man experts do these gold annotations. Finally, every essay has a resolved score. This setting is similar to standardized English examination patterns like TOEFL, SATs. The standard examination involves multiple human grading of an essay for a robust and reasonable final score. Most of these essays have an average word length of 550 words with every set representing essays with distinctive traits. There are around 12,978 essays across the eight sets of essays, and the split of 80:20 is followed for training and testing on every split. This leads to a comprehensive way to score individual essays in the dataset.

The hardware used is CPU: Intel(R) Xeon(R) L5640 @2.27 GHz 2.26 GHz; RAM: 8G; HDD: 100 GB; GPU: GTX 1080i. The software environment is under Windows 10, Python 3.6, TensorFlow—gpu 1.4.

Evaluation metrics

In this work, we computed the Quadratic Weighted Kappa (QWK) score (Vaswani et al., 2017) for evaluating the proposed model. The QWK score is considered as a robust metric analysis performance of essay scoring systems. Its values lie between 0 (no agreement) and 1 (complete agreement). Table 1 shows ASAP statistics, essays have been divided into 8 different categories which is indicated by Prompts sets in Table 1, Prompt set represents the collection of essay prompts. The model uses these essay prompts to construct the remaining part of the essay. Here, we calculated QWK score between the automated essay scores and the resolved scores for

Table 1 Statistics of the ASAP dataset (Uzun, 2018)

Prompts set	Essays	The average length of essays	Score range
1	1783	350	2–12
2	1800	350	1–6
3	1726	150	0–3
4	1772	150	0–3
5	1805	150	0–4
6	1800	150	0–4
7	1569	250	0–30
8	723	650	0–60

every essay set separately. In order to calculate QWK, the weights are assigned to every cell of the contingency table. These assigned weights range from the value 0 to the value 1. Weight value of 1 is assigned to the diagonal cells. The proportion of observed agreement Oa can be described as the calculated summation of weighted proportions using Eq. 4.

$$O_a = \sum_m \sum_n T_{mn} K_{mn} \tag{4}$$

The proportion of expected chance agreement Ea can be described as the collected summation of weighted multiplication of rows and columns and can be calculated as per Eq. 5.

$$E_a = \sum_m \sum_n T_{mn} K_m + K + n \tag{5}$$

QWK score between Ea and Oa is computed using Eq. 6.

$$QWK = \frac{O_a - E_a}{1 - E_a} \tag{6}$$

The mean value of QWK score is calculated using obtained kappa score values through Fischer Transformation (Bond & Richardson, 2004). Here, it is hypothesized that a well-written essay must have coherency in its sentences. The perfect way to ensure coherency is computing the similarity score between a sentence and its neighbouring sentences with the idea being that closer sentences will have more coherence than far away sentences.

To this end, we followed the technique proposed in Nadeem et al. (2019). Firstly, the similarity of the sentence with its neighbouring sentence is calculated and represented as Sim2. Secondly, the similarity between the sentences with all other sentences is also measured and represented as Simall. In this work, we used cosine similarity function as most commonly used distance measurement vector. We extracted the embeddings from the last hidden state of the LSTM, present on top of RoBERTa's architecture for computing similarity. It is also worth noting that this embedding is used to perform the regression-based task.

Apart from QWK, we also use f1 score to evaluate the results of our model. F1 score is a good metric to know the accuracy of the system for each score from 0 to 50. We find out that our model is giving above 78% f1 score against all the sets which is a great result. Compared to other models, our method has performed better for accurately predicting the true score of the essay with respect to the human ratings. Table shows the results in terms of f1 score for all the eight essay sets.

Results and discussion

The experimental results obtained in this set of experiments are presented in Table 2. The obtained values represent average QWK score between the proposed model for automated essay scoring system (RoBERTa + Bi-LSTM) and two human raters

Table 2 Comparative experimental results of the proposed model and human raters

Des	PS1	PS2	PS3	PS4	PS5	PS6	PS7	PS8	Avg. QWK
AES/h1	0.83	0.74	0.76	0.85	0.79	0.87	0.79	0.72	0.791
AES/h2	0.85	0.71	0.71	0.84	0.86	0.88	0.82	0.64	0.782

denoted by h1 and h2 as a baseline for comparison purpose. The closeness between the average QWK scores for two different human raters shows the coherency level achieved by our proposed model. Essay differ in domain but the structure of the language is same, so giving similar results for all the sets shows that our model is performing better in terms of coherency. AES/h1 is an experimental setting where the AES system is measured against human rater, h1. It is similar to the case of AES/h2. The closeness of the AES/h1 and AES/h2 denotes the fact that it is robust to human rater change.

Figure 6 presents the variation in training loss with epochs for the proposed model in comparison to state of the art models. In this work, we considered TSLF (Nadeem et al., 2019), a feature-based baseline (that includes dependency parsing features, POS tagging features), BERT-HAN, NLI-DCM-BCA and NLI-DCM-BCA + features (Nadeem et al., 2019) based models for comparative analysis. It can be observed from Fig. 6 that the proposed model (RoBERTa + Bi-LSTM) exhibits the best convergence pathway as compared to its counterparts.

To validate that our model is not over-fitting, we will also plot the test loss curves for our model and the other models. Loss curves in both training and test justifies the choice of choosing 180 training epochs. As our model converges very fast and after that it does not decay much (Fig. 7). Looking at the test curve shows that over model is not over-fitting at all. This is because of the embedding space adjustment for essay domain in Roberta fine tuning which we performed before training the model on ASAP task (Fig. 8).

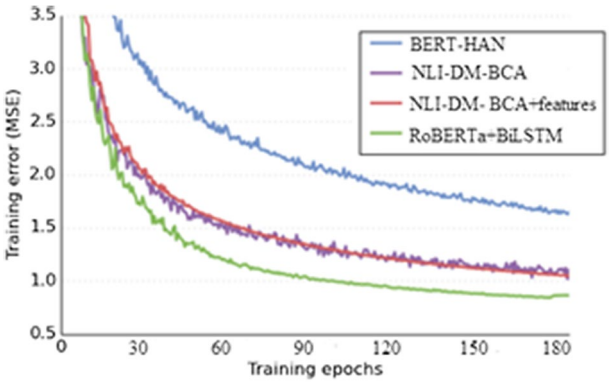

Fig. 6 Training loss of various models

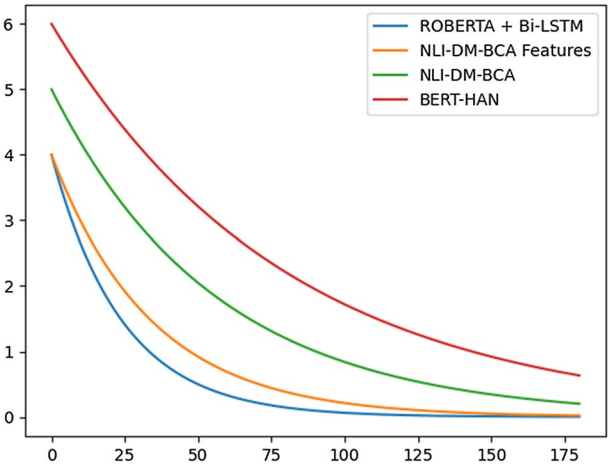

Fig. 7 Test Loss of different models

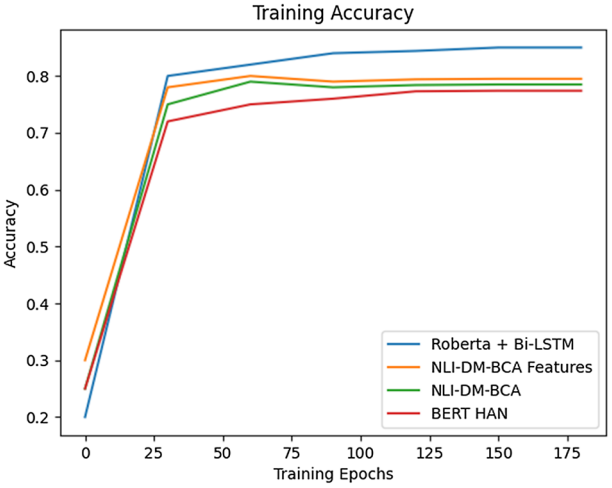

Fig. 8 Training Accuracy Comparison

To check the model in terms of accuracy, we plot accuracy of our model and compare it with the other models so that we can validate our findings. Figure 9 shows the training accuracy of our model with the other s models for essay scoring. Our model outperforms NDI and simple BERT based models by a wide margin.

Our model does not overfit and to prove that we trained the other models for 180 epochs and they all resulted in over-fitting. This is because we have fine tuned the model on essay dataset which prevents the model from over-fitting when trained on ASAP dataset. Validation of accuracy of our model was almost

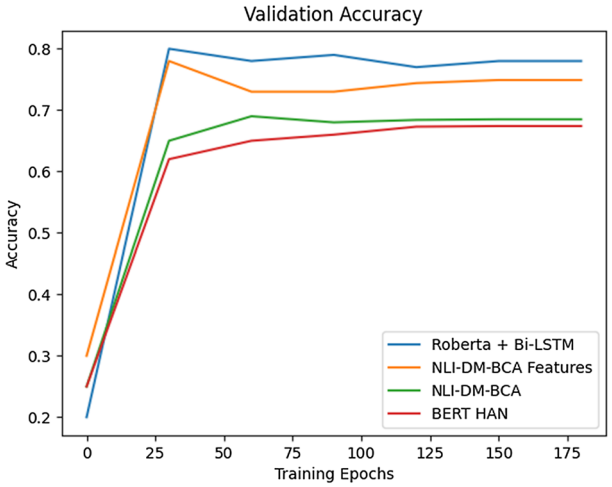

Fig. 9 Validation Accuracy Comparison

on par with the training accuracy but in other cases the validation accuracy was very low compared to the training accuracy when those models.

Table 3 shows numerical results for the proposed model in comparison to the other state of art models consisting of deep learning-based, feature-based, and ensemble models. It can be observed from Table 3 that the proposed model outperforms the existing models on the ASAP dataset in terms of QWK score.

It can be concluded from the above-cited experimental results that the proposed model outperforms human raters and state of the art models consisting of deep learning-based, feature-based, and ensemble models in terms of QWK score and performance convergence. This demonstrates the validity and superiority of the proposed model in comparison to the other models.

Table 3 Comparative experimental results of the proposed model and state of the art models

Model used	Average QWK
TSLF	0.76
Feature-based baseline	0.74
BERT-HAN	0.68
NLI-DM-BCA	0.73
NLI-DM-BCA + Features	0.77
RoBERTa + BLSTM (Ours)	0.80
BLSTM	0.66

Conclusion

Advancements in AI technology have made automated grading of essays a realistic option. The state-of-the-art research highlights the potential of AES to be used widely in higher educational assessments in the near future. However, several shortcomings have been reported in the literature that limits the deployment of AES systems in the higher education system, which indicate the need of improving these systems for attaining results better than the human raters. The present study is also an effort to address certain limitations present in the existing AES systems. This work demonstrated a transformer-based neural network model for improved performance on automated essay scoring based on the Kaggle's ASAP dataset. The proposed model addressed the issue of coherency in essays that are ignored by traditional essay score methods, including traditional NLP pipelines, deep learning-based methods, a mixture of both. The proposed model used Bi-LSTM model on the top of pre-trained RoBERTa language model, a state- of-the-art deep learning technique to address the issue of coherency in essays by keeping other features intact. The performance of the proposed model is computed on ASAP dataset for automated essay scoring. The experimental results on essay scoring are compared with human raters and state of the art in the field. It is concluded from results that the proposed model outperforms the existing methods in essay scoring. The comparative analysis of experimental results demonstrates the applicability and superiority of the proposed model in the context of automated essay scoring at higher education level. In future, we may focus on ways to solve the problem of bad-faith essay submissions in an automated fashion, possibly by researching the genre of generative methods for automated essay generation.

Note: This work is an extension of the paper titled "Essay Scoring Tool by Employing RoBERTa architecture" submitted at the International Conference on Data Science, E-learning and Information Systems 2021.

References

Alikaniotis, D., Yannakoudakis, H., & Rei, M. (2016). Automatic text scoring using neural networks. arXiv preprint arXiv:1606.04289.

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

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

Bansal, M., & Passonneau, R. J. (2018). Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: Tutorial abstracts. In *Proceedings of the 2018 conference of the north american chapter of the association for computational linguistics: Tutorial abstracts* (2018).

Bennett, R. E., & Bejar, I. I. (1998). Validity and automad scoring: It's not only the scoring. *Educational Measurement: Issues and Practice, 17*(4), 9–17.

Bernstein, J., Van Moere, A., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing, 27*(3), 355–377.

Beseiso, M., & Alzahrani, S. (2020). An empirical analysis of bert embedding for automated essay scoring. *International Journal of Advanced Computer Science and Applications*. https://doi.org/10.14569/IJACSA.2020.0111027.

Bond, C. F., & Richardson, K. (2004). Seeing the fisherz-transformation. *Psychometrika, 69*(2), 291–303.

Cer, D., Yang, Y., Kong, S. Y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., et al. (2018). Universal sentence encoder. arXiv preprint arXiv:1803.11175.

Citawan, R. S., Mawardi, V. C., & Mulyawan, B. (2018). Automatic essay scoring in e-learning system using lsa method with n-gram feature for Bahasa Indonesia. In *MATEC web of conferences*, vol. 164, p. 01037. EDP Science.

Cummins, R., & Rei, M. (2018). Neural multi-task learning in automated assessment. arXiv preprint arXiv:1801.06830.

Cushing Weigle, S. (2010). Validation of automated scores of TOEFL iBT tasks against non-test indicators of writing ability. *Language Testing, 27*(3), 335–353.

Dascalu, M., Dessus, P., Bianco, M., Trausan-Matu, S., & Nardy, A. (2014). Mining texts, learner productions and strategies with readerbench. In *Educational data mining*, pp. 345–377. Springer.

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.

Dong, F., Zhang, Y., & Yang, J. (2017). Attention-based recurrent convolutional neural network for automatic essay scoring. In *Proceedings of the 21st conference on computational natural language learning (CoNLL 2017)*, pp. 153–162.

Dos Santos, C., & Gatti, M. (2014). Deep convolutional neural networks for sentiment analysis of short texts. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers*, pp. 69–78.

Farag, Y., Yannakoudakis, H., & Briscoe, T. (2018). Neural automated essay scoring and coherence modeling for adversarially crafted input. arXiv preprint arXiv:1804.06898.

Foltz, P. W., Laham, D., & Landauer, T. K. (1999). Automated essay scoring: Applications to educational technology. In *EdMedia+ innovate learning*, pp. 939–944. Association for the Advancement of Computing in Education (AACE).

Gierl, M. J., Latifi, S., Lai, H., Boulais, A. P., & De Champlain, A. (2014). Automated essay scoringand the future of educational assessment in medical education. *Medical Education, 48*(10), 950–962.

Ginther, A., Dimova, S., & Yang, R. (2010). Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implications for automated scoring. *Language Testing, 27*(3), 379–399.

Hartley, D. J. (2004). Automated language and interface independent software testing tool (2004). US Patent 6,763,360.

Hazar, M. J., Toman, Z. H., & Toman, S. H. (2019). Automated scoring for essay questions in e-learning. In *Journal of Physics: Conference Series*, vol. 1294, p. 042014. IOP Publishing.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Journal of Neural Computing, 9*(8), 1735–1780.

Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional lstm-crf models for sequence tagging. arXiv preprint arXiv:1508.01991.

Hussein, M. A., Hassan, H., & Nassef, M. (2019). Automated language essay scoring systems: A literature review. *PeerJ Computer Science, 5*, e208.

Kwong, A., Muzamal, J. H., & Khan, U. G. (2019). Automated language scoring system by employing neural network approaches. In *2019 15th international conference on emerging technologies (ICET)*, pp. 1–6. IEEE.

Lample, G., & Conneau, A. (2019). Cross-lingual language model pretraining. arXiv preprint arXiv:1901.07291.

Larkey, L. S. (1998). Automatic essay grading using text categorization techniques. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 90–95.

Latifi, S., Gierl, M. J., Boulais, A. P., & De Champlain, A. F. (2016). Using automated scoring to evaluate written responses in English and French on a high-stakes clinical competency examination. *Evaluation & the Health Professions, 39*(1), 100–113.

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In *International conference on machine learning*, pp. 1188–1196.

Li, D., Zhong, S., Song, Z., & Guo, Y. (2020). Computer-aided English education in china: An online automatic essay scoring system. In *International conference on innovative mobile and internet services in ubiquitous computing*, pp. 264–278. Springer.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Machicao, J. C. (2019). Higher education challenge characterization to implement automated essay scoring model for universities with a current traditional learning evaluation system. In *International conference on information technology & systems*, pp. 835–844. Springer.

Mahlangu, V. P. (2018). The good, the bad, and the ugly of distance learning in higher education. Trends in E-learning pp. 17–29.

Nadeem, F., Nguyen, H., Liu, Y., & Ostendorf, M. (2019). Automated essay scoring with discourse-aware neural models. In *Proceedings of the fourteenth workshop on innovative use of NLP for building educational applications*, pp. 484–493.

Nadeem, F., & Ostendorf, M. (2018). Estimating linguistic complexity for science texts. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, pp. 45–55.

Ng, S. Y., Bong, C. H., Hong, K. S., & Lee, N. K. (2019). Developing an automated essay scorer with feedback (aesf) for Malaysian university English test (muet): A design-based research approach. *Pertanika Journal of Social Sciences & Humanities, 27*(2).

Page, E. B. (1967). Grading essays by computer: Progress report. In *Proceedings of the invitational conference on testing problems.*

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543.

Perin, D., & Lauterbach, M. (2018). Assessing text-based writing of low-skilled college students. *International Journal of Artificial Intelligence in Education, 28*(1), 56–78.

Phandi, P., Chai, K. M. A., & Ng, H. T. (2015). Flexible domain adaptation for automated essay scoring using correlated linear regression. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pp. 431–439.

Ramineni, C., Trapani, C. S., Williamson, D. M., Davey, T., & Bridgeman, B. (2012). Evaluation of the e-rater R scoring engine for the gre R issue and argument prompts. *ETS Research Report Series, 2012*(1), i–106.

Reilly, E. D., Williams, K. M., Stafford, R. E., Corliss, S. B., Walkow, J. C., & Kidwell, D. K. (2016). Global times call for global measures: Investigating automated essay scoring in linguistically-diverse moocs. *Online Learning, 20*(2), 217–229.

Rong, X. (2014). word2vec parameter learning explained. arXiv preprint arXiv:1411.2738.

Rudner, L. M., & Liang, T. (2002). Automated essay scoring using Bayes' theorem. *The Journal of Technology, Learning and Assessment*, *1*(2).

Shermis, M. D., Hamner, B. (2013). 19 Contrasting state-of-the-art automated scoring of essays. *Handbook of automated essay evaluation: Current applications and new directions*, p. 313.

Shi, W., & Demberg, V. (2019). Next sentence prediction helps implicit discourse relation classification within and across domains. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 5794–5800.

Su, M. H., Wu, C. H., & Zheng, Y. T. (2016). Exploiting turn-taking temporal evolution for personality trait perception in dyadic conversations. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 24*(4), 733–744.

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.

Tay, Y., Phan, M. C., Tuan, L. A., & Hui, S. C. (2017). Skipflow: Incorporating neural coherence features for end-to-end automatic text scoring. arXiv preprint arXiv:1711.04981.

Tobback, E., Naudts, H., Daelemans, W., de Fortuny, E. J., & Martens, D. (2018). Belgian economic policy uncertainty index: Improvement through text mining. *International Journal of Forecasting, 34*(2), 355–365.

Uzun, K. (2018). Home-grown automated essay scoring in the literature classroom: A solution for managing the crowd? *Contemporary Educational Technology, 9*(4), 423–436.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008.

Vaughn, D., & Justice, D. (2015). On the direct maximization of quadratic weighted kappa. arXiv preprint arXiv:1509.07107.

Wang, Y., Wei, Z., Zhou, Y., & Huang, X. J. (2018). Automatic essay scoring incorporating rating schema via reinforcement learning. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 791–797.

Wirth, C., & Fürnkranz, J. (2014). On learning from game annotations. *IEEE Transactions on Computational Intelligence and AI in Games, 7*(3), 304–316.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pp. 5753–5763.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016). Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pp. 1480–1489.

Yin, W., Ebert, S., & Schütze, H. (2016). Attention-based convolutional neural network for machine comprehension. arXiv preprint arXiv:1602.04341.

Yu, W., & Barker, T. (2020). A study on the effectiveness of automated essay marking in the context of a blended learning course design. *Education Language and Sociology Research, 1*(1), 20.

Zechner, K., Higgins, D., Xi, X., & Williamson, D. M. (2009). Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication, 51*(10), 883–895.

Zhang, H., Magooda, A., Litman, D., Correnti, R., Wang, E., Matsmura, L., Howe, E., & Quintana, R. (2019). erevise: Using natural language processing to provide formative feedback on text evidence usage in student writing. In *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, pp. 9619–9625.

Zhang, M. (2013). Contrasting automated and human scoring of essays. *R & D Connections, 21*(2), 1–11.

Zhang, Y., & Wallace, B. (2015). A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. arXiv preprint arXiv:1510.03820.

Zhu, W. (2019). A study on the application of automated essay scoring in college English writing based on PIGAI. In *2019 5th International conference on social science and higher education (ICSSHE 2019)*, pp. 451–454. Atlantis Press.

Dr. Majdi Beseiso is an assistant professor, Computer Science department, at Al-Balqa Applied University, Jordan. He received his B.Sc. degrees in computer science from Jordan University of Science and Technology in 2003 and received his M.Sc. degree in computer science from University of Jordan in 2006. In 2013 he obtained his Ph.D. degree in Information and Communication Technology from Tenaga National University, Malaysia. Dr. Beseiso research interests include Semantic Web and Web Content Mining, Information Extraction and Information Retrieval from Web, NLP, Artificial Intelligence, machine learning, Software Engineering & E-Learning. Dr.Beseiso is the dean assistant for planning, development and quality at Prince Abdullah Bin Ghazi Faculty of Information and Communication Technology.

Dr. Omar A. Alzubi received his B.Sc. degree from University of Jordan, Jordan. He obtained his M.Sc. degree with distinction in Computer and Network Security from New York Institute of Technology, New York, USA in 2006. In 2013, he received his Ph.D. degree in Computer and Network Security from Swansea University, Swansea, UK. Currently he is an associate professor in the Computer Science Department at Al-Balqa Applied University, Jordan. Dr. Alzubi research interests include computer and network security, machine learning, and cryptography. His cumulative research experience for over ten years resulted in publishing more than thirty articles in highly impacted journals. Dr. Alzubi is the vice dean of Prince Abdullah Bin Ghazi Faculty of Information and Communication Technology. In addition he is an editorial board member and reviewer of many prestigious journals in computer science field.

Dr. Hasan Rashaideh is an associate professor, Computer Science department, at Al-Balqa Applied University, Jordan. He received his B.Sc. and M.Sc. degrees in computer science and information technology from Yarmouk University in 1999 and 2002 respectively. In 2008 he obtained his Ph.D. degree in computer science from Saint Petersburg Electrotechnical State University-Russian Federation. Then, he joined the computer science department at Prince Abdullah Bin Ghazi Faculty of Information and Communication Technology as an assistant professor. Dr. Rashaideh was appointed as Head of the department from Jul. 2015 to July 2018. His research interests include machine learning, image processing, and computer vision, information retrieval, and optimization.