Journal of Theoretical and Applied Information Technology
31st March 2021. Vol.99. No 6

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

AEGD: ARABIC ESSAY GRADING DATASET FOR MACHINE LEARNING

BASSAM AL-SHARGABI [1] , RAWAN ALZYADAT [2], AND FADI HAMAD[3]

[1,2] Faculty of Information Technology, Middle East University, Amman-Jordan,

[3] Faculty of Information Technology, Isra University, Amman -Jordan,

E-mail: [1]bshargabi@meu.edu.jo., [2]alzyadtrawan@gmail.com, [3]F.hamad@iu.edu.jo

ABSTRACT

Recently, developing an Automatic Essays Grading (AEG) system has become an attractive topic in industry and academia. Most of the grading systems rely on machine learning to grade the essays based on a predetermined dataset. However, English essays scored based on Automated Student Assessment Prize (ASAP) dataset whereas the absence of such a dataset for Arabic essays is a major predicament. Therefore, in this paper, we have established the Arabic Essay Grading Dataset (AEGD) that is suitable for machine learning to develop an Arabic AEG system. This dataset comprises a collection of essay questions along with its graded model answers for several topics that cover various school levels. We used the Naive Bayes (NB), Decision tree (J48), and meta classifier as a well-known machine learning algorithms to evaluate and test the established AEGD. The results show that the accuracy rates of the three classifiers have reached 79%, 81%, and 86% based on the established AEGD..

Keywords: *Automated Essay Grading; Arabic Essay Grading; Dataset; Machine Learning; Classification Algorithm*

1. INTRODUCTION

Recently, the Automatic Essay Grading (AEG) system plays a vital role because it serves colleges and schools to improve and speed up the process of essays grading and to avoid human errors or any kind of bias that might affect the integrity of the essay grading process. AEG development draws the attention of schools, universities, and researchers, where various research has taken place in literature to develop more accurate and effective methods to automate essay grading. Moreover, the margin of errors between AEG system and human grading have almost reached an acceptance level [1]. Currently, most of the AEG systems are based on the use of Artificial intelligence such as machine learning and deep learning as an effective method for developing an AEG system. The use of machine learning to grade essays in the context of a classification problem, where essays are graded based on the existence of a predetermined set of graded model answers or a huge dataset of questions with a human graded model answers as class labels. Novel approaches in machine learning and deep learning have revealed that utilizing neural network techniques for AEG has achieved more efficient outcomes [2]. Furthermore, Most of the AEG systems that designed based on machine learning and deep learning relies on extracted features from all essays graded model answers that are automatically learned from a large dataset such as the ones we are proposing in this paper.

The scope of this paper falls into developing an AEG system, where most of the AEG employs Natural Language Processing (NLP) concepts and machine learning methods to automate the process of grading the written essays. The NLP and Machine learning methods rely on extracting features from the written essays of students and training essays dataset using a neural network to tackle the predicament of automating the essay grading. The feature extracting approaches can easily help to foretell grades using a set of features such as essay length or spelling errors. These features also can be adaptable to be changed based on set criteria to grade essays based on human feedback.

The problem that we are dealing with in this paper is the non-existence of such dataset for Arabic essays. Whereas in English the existence of

Automated Student Assessment Prize (ASAP) dataset , where it contains over more than 13,000 essay in English. The ASAP dataset utilized to enhance and improve designing a machine learning based AEG system for grading English essays [3,4,5]. For Arabic essays, there are few attempts in literature. These attempts end up establishing a small and insufficient dataset that the researchers can use or exploit to develop an AEG system for Arabic written Essays. The objective of this paper is to establish a dataset for Arabic essay grading that can be used to build models using machine learning to grade or score students essays in schools or universities. This dataset can also be used by researchers in field of machine learning and Natural Language Processing (NLP) to develop effective methods and approaches for grading Arabic essays.

Establishing Arabic Essays Grading Dataset (AEGD) requires gathering essay questions with its model answers from reliable sources, and a good selection strategy. Also, when collecting the dataset it is important to consider the precision of the collected data. In this paper, the collected AEGD has taken from the teacher's guide of Islamic, History, Geography, Biology, Computer, Geology, Chemistry, and Physics. These subjects are taught in Jordanian schools from 9th to 11th grades. The collected data were analyzed and structured to ensure that the dataset is suitable for machine learning.

The rest of the paper organized as follows: Section 2 outlines topics background and the most recent related work. The elaboration of the methodology of the proposed Datset in section 3 experimental result and discussion are presented in section 4. Finally, the conclusion about this paper is drawn in Section 5.

2. BACKGROUND AND RELATED WORK

The background of essay grading systems, machine learning, and most recent related works are described and discussed in this section.

2.1 Automated Grading System

AEG systems used to grade students' essays without human interference, where students' essays are graded or evaluated to overcome time, cost, and human bias or errors. AEG systems were developed with the intent of helping teachers and college professors in grading student essays in an effective manner using reliable methods [1]. In the literature there are many types of AEG systems, the first system is the Project Essay Grader (PEG) is established by Ellis Page in 1966. The essay grade in PEG is estimated based on the essay quality by the regression model and certain features such as word length, paper length, conjunction, and punctuation marks[4]. The second AEG system is the Intelligent Essay Assessor (IEA) introduced by Landauer and Foltz in 1997 [5]. The essays scoring by IEA by measuring the similarities between students' essays and model essays. The similarities can be determined by the most frequent word in both essays. The scoring of IEA system revealed a high correlation with the human scores. Finally, the iElectronic Essay Rater (E-rater) uses certain features to determine the similarity between student essay and model essay [6].

Recently, most AEG system were developed with more efficient approaches using machine learning and deep learning. For example, written essays were graded for L2 learners of English, the grading was identified using a set of features that were set to examine the proficiency level of the learners[7] . They used these written essays to grade new students essays based on machine learning methods for feature extraction. The extracted feature from both existing graded essays in a dataset with the new essay are compared to determine which of existing essay is more similar to the new essay to be graded [3].

Different essay grading systems used machine learning to predict essay grades using different machine learning algorithms such as Support Vector Machine(SVM), Naive Bayes(NB), Random Forest. However, the linear regression proved that the machine learning algorithms can provide a comparable accuracy regarding predicting essay grades when compared to human graders. As the linear regression result showed the ability to predict a student's score it is good as shown in. They used a dataset from kaggel.com, by William and Hewlett that contains 1300 English essays. The essays contain almost 150 to 550 words in length. The essay dataset was divided into 8[8]. Furthermore, the deep learning approaches were also introduced but it involves using a huge dataset of essays already been graded by human graders as mentioned earlier to be used in the phase of training. The Long Short Term Memory (LSTM) is a popular deep learning approach used in text classification also been used to grade students' essays. The LSTM estimates the grades of essays, where the essays were processed as a vector containing a list of words for each essay. Moreover, a Convolutional Neural

Network(CNN) and LSTM system also proposed to predict student essay grades in [9], which proved to enhance the accuracy of grade prediction when compared to LSTM.

2.2 Machine Learning

Basically, machine learning is the utilization of artificial intelligence (AI) techniques to provide a system with the capabilities of learning and improving from past experiences without explicit programming. Machine learning also can be defined as a way to design algorithms that can obtain input data, and exploit a set of statistical and mathematical analysis to predict outputs. [10]. The four common types of machine learning are supervised, unsupervised, semi-supervised, and reinforcement learning.

The first type of machine learning is the supervised learning approach that works as a mapping function that uses set training labeled data that used to predicts the outcome of new data. This approach provides an outcome for any new input data after good training examples if it improperly is predicted. A comparison is made between the new output data and the correct outcome and try to discover the error and correct the model. The second type of machine learning is unsupervised learning that deals with grouping unlabelled data into similar groups without guidance. The third type is semi-supervised learning which is a mixture of the supervised and unsupervised machine learning. During the training in a semi-supervised approach, it combines a set of labeled data with an enormous amount of unlabeled data. Finally, the reinforcement approach relies on predicting the outcome of the new data through a self-learning process with its environment [11].

Using Machine learning to grade written essays involves four main phases, Preprocessing essays text inclucding Tokenization, stop-words removal, stemming or lemmatization). Then feature extraction phase where the most popualr method for extractiong featre are Term Frequency ,Term Frequency-Inverse Document Frequency (TF-IDF), Bag of words followed by text representation , and last phase is the text classification. The most critical phase is feature extraction in machine learning and plays a significant role in essay classification, where it minimizes the computational overhead and enhances the classifier's accuracy to predict the correct class label. Therefore, researchers in various disciplines have become aware of the importance of feature extraction and that resulted in proposing novel and enhanced feature extraction methods in machine learning [12]. Predicting the grade of any essay can be treated as a classification problem but the high dimensionality nature of text datasets can be tricky. Therefore, a good feature extraction method with ensemble machine learning can help to improve the classification performance [13].

In text classification, the datasets contain a huge number of unique words and it depends on the nature of the dataset. The unique words can be determined as an output of the preprocessing phase of machine learning. This results in huge dimensional datasets but this problem can be tackled using the dimensionality reduction in preprocessing phase which will lead to an efficient and enhanced performance of the classifiers [12].

In this paper, we have tested and validated the Arabic essay grading data set using Naïve Bayes (NB), Decision Tree (J48), and Meta- classifier. The first classifier is the NB, where it has been widely used in text classification [14], where the text is seen as a set of words and the sequence of words in the text irrelevant. The NB classifier is based on a set of probabilities to determine which text belongs to the given class labels. Regarding the NB simplest form, the process of determining which text belongs to a class label is based on counting the frequency of a word in texts using TFIDF as a feature extraction method. Therefore, the NB is very simple and efficient for the process of text classification and also provides good performance when compared to other classifiers because it can be used with fewer data in the training phase. The NB is mainly based on the solid hypothesis of the data distribution [15], [16]. The basic problem when using NB, that it is based on the assumption that features in feature space are independents which might lead to inadequate performance of the classifier.

The second classifier is the J48 Decision tree, which is used widely in different domains for classification. The decision tree relies on splitting the dataset into smaller subsets based on a predefined attribute and places it in the tree branches for simplicity and organizational purpose. The J48 is an extended version of C 4.5 (ID3), where the J48 classifier capable of treating issues such as missing value, and trees pruning to tackle the over-fitting problem. J48 gain its popularity because it is a very simple and fast classifier in training and predictions.

The third classfier is Meta classifier, which is an ensemble learning technique to mix various classifiers via a meta-classifier. Each classifier is

trained and based on a subset dataset and their output (predictions) are stacked and used as meta-features to train the meta classifier which delivers the final prediction. Therefore, the meta classifier uses classes predicted by several classifiers and chooses the final one that achieves a better accuracy rate.

In this paper, the focus only on supervised machine learning mainly the classification algorithms. The classification algorithms main goal is to predict the data label based on learning from past experiences, i.e. training data. The second type is a regression which also includes predicting data label but for numerical training data[17]. In this paper, we only concentrated on the application of classification algorithms to predict students essay grades based on a set of training example of written essays as model answers already graded, where the grade here is the class label.

2.3 Related Work

A text similarity approach was used to introduced to grade student Arabic essay questions[18]. The questions were broken into two groups based on the length of the model answer as a long and short answer. They used a dataset that comprises 21 questions with 210 short model answers. Their system was based on text similarity to grade student essay questions and then compared the results with the model answer. The text-similarity measures applied to the extracted features from both student answer and model answers to determine the score of the student's answer based. They also used the Arabic WordNet tool to generate terms that might have the same meanings or contexts, where it is prepared using the notion of a synset.

An automated Arabic essay grading model proposed by employing the support vector machine (SVM) with text similarity algorithms [19]. The model grades essays based on F-score to extract features from students' essays and typical answers along with the use of the Arabic WordNet (AWN) as an important knowledge-based method for semantic context. The goal of applying the AWN is to discover all relevant words from student answers to provide all words synonymous that might relate to the typical answers and produce a fair score for the student. They used a dataset that contains 40 questions and 120 answers to test their model.

An automated scoring system for Arabic essays was based on translating the text of the essay from Arabic to English because of the shortage or pre-processing of Arabic language texts such as (stemming and lemmatization) [20]. They used a small dataset that contains only 610 short answers which have been translated from Arabic into English. The dataset included four classes of questions in the form of define, what, why, and explain. They used the k-mean clustering method and similarity measures to score student answers.

Another AEG system based in Latent Semantic Analysis (LSA) is used to grade student's essays written in the Arabic language [21]. The LSA was exploited to extract semantic and syntactic features. The syntactic feature makes the accuracy of AEGS more effective. They relied on a dataset that contains only 61 questions and each question has 10 typical graded answers. To validate their methods, Arabic essays were graded using Term Frequency-Inverse Document Frequency (TF-IDF) and LSA, the outcome by using LSA is 0.745 compared to other methods. Accordingly, the number of essay questions were just limited to 61, and the results could not be generalized unless the methods were applied to more comprehensive dataset such as the one we proposed in this paper. Another approach based on LSA and Cosine Similarity measure was used, where it relied on enhancing LSA matrix (WCM) through using proper methods for pre-processing Arabic text such as unifying the form of letters, removing the formatting, substituting synonyms, stemming and using more comprehensive list of Arabic list of stop words to be removed in order to produce a matrix that depicts Arabic texts in a more efficient way compared to the conventional LSA matrix [22]. Moreover, a method based on LSA with Rhetorical Structure Theory was produced in [23]. The method evaluated written essays and scored by individual teachers for grades from 7-12. The number of essays was only 350 and their evaluation was based on language proficiency, and the structure of the essay.

3. METHODOLOGY

In this section, the proposed methodology that was adopted to establish the AEGD for machine learning is illustrated in Figure 1. methodology includes, Dataset collection and structure, Data pre-processing, Training dataset, and Validation and evaluation of dataset.

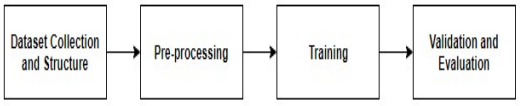

Figure 1: Methodology

Journal of Theoretical and Applied Information Technology
31st March 2021. Vol.99. No 6

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

3.1 Data Collection and Structure

The dataset collected from various subjects taught across different Jordanian schools on different levels such as Islamic, History, Geography, Biology, Computer, Geology, Chemistry, and Physics. The goal of collecting the dataset form a various topic or subjects was to ensure that our dataset contains a vast number of unique words or vocabulary in Arabic. The aggregated dataset comprises a collection of questions and model answers graded by teachers and validated by education experts and professors. the collected AEGD has taken from the teacher's guide of Islamic, History, Geography, Biology, Computer, Geology, Chemistry, and Physics. These subjects are taught in Jordanian schools from 9th to 11th grades. For each question there was only one model answer in teachers book, but after collecting the dataset we got two additional model answers based on teacher's feedback. The established dataset as shown in table 1 contains 1003 questions and 3009 answers, for 8 subjects to exhibit the diversity of words as illustrated in figure 2. The collected dataset contains 35615 words and 10364 unique words and we have conducted grammar and spelling checks during the collection of such data.

Table 1: Collected Dataset description

Topic	***Number of questions***	***Number of answers***	***Number of words***	***Number of unique words***
Islamic 12	213	639	7192	1050
Islamic 11	109	327	2241	774
Geography 12	107	321	4499	1307
Geography 11	74	222	3261	1118
History	190	570	3541	1229
Biology	80	240	3779	1361
Computer	88	264	4257	1578
Geology	81	243	4747	1140
Chemistry	35	105	1186	450
Physics	26	78	912	357
Total	1003	3009	35615	10364

The established dataset was structured and designed to be suitable for machine learning algorithms. Therefore. each subject of the collected data was designed with four main attributes as shown in table 2 (Sample of two questions of the dataset for two subjects in Table 2, the first question was taken from Islamic 12 subject, and second question was taken from Geography subject). The attributes are Essay_id, Essay_Question, Answer, and Grade as described as follows:

- Essay_id: question number.
- Essay_qustion question text, where the questions vary in length as the number of words ranges (3-25 words).
- Answer: Answer: with each essay question there are 3 model answers, where each answer is different in length as the number of words ranges (0-100 words).
- Grade: the three model answers were graded by human as follows 5,2, and 0.

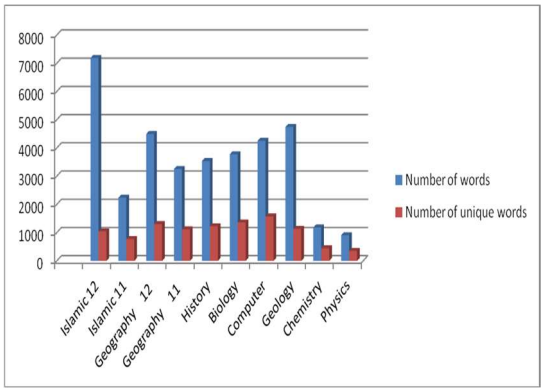

Figure 2: Number of words or terms in each Subject

The Grade is the target class attribute, the values (5,2,0) of the target class outline the grades of the essay as established by a human grader. The dataset was structured and format based on using a comma-separated value (CSV) and Attribute-Relation File Format (ARFF)file, where the file involves a set of instances with a set of attributes. These instances are the essays model answer, and grade. Each instance represents one essay model answer. We used WEKA ARFF file in our experiments while the CSV files can be used for any tools of machine learning.

3.2 Dataset Pre-Processing

The pre-processing is a challenging task for the essay questions written in Arabic, but this task is important because it omits unnecessary words; for example, eliminate stop words, eliminate conjunctions. The major tasks of dataset pre-processing are the following:

- Tokenization: Is the first step, where texts such as essay questions and answers were divided into chunks called terms. The terms are separated by punctuation such as comma, and space, regardless of the meaning or relation of these words.

Table2: Sample of Essay Questions and graded model answer

Essay_id	*Essay-question*	*Answer*	*Grade*
1	ما المقصود بالقيم السياسية في الاسلام What is meant by political values in Islam	هي المبادئ والقواعد المستمدة من القرآن الكريم والسنة النبوية التي تضبط علاقة مؤسسات الحكم بعضها ببعض، وعلاقتها بالأفراد، وتضبط علاقة الدول بغيرها من الدول. They are the principles and rules derived from the Noble Qur'an and the Sunnah of the Prophet that govern the relationship of institutions of government with one another, their relationship with individuals, and control the relationship of states to other countries.	5
		هي المبادئ والقواعد المستمدة من القرآن الكريم والسنة النبوية. They are the principles and rules derived from the Noble Qur'an and the Sunnah of the Prophet.	2
		فارغ. Empty	0
2	ما المقصود بمسقط الخريطة What is the map projection?	نقل المعلومات والبيانات من سطح الأرض الكروي الى سطح مستوي على الخريطة، بحيث تبقى معالم سطح الأرض في مواقعها الصحيحة بالنسبة الى بعضها بعضاً وفق معادلات رياضية. Transferring information and data from the spherical Earth's surface to a flat surface on the map, so that the features of the Earth's surface remain in their correct locations relative to each other according to mathematical equations.	5
		نقل المعلومات والبيانات من سطح الأرض الكروي الى سطح مستوي. Transferring information and data from the spherical Earth's surface to a flat surface on the map.	2
		فارغ. Empty	0

- Normalization: Normalization is important for processing Arabic text, where text words were transformed into a standard form such as eliminating diacritics (where the term العَدُوّ converted to العدو) , and elongation from input words (where the term يـــســـاعـــد converted to يساعد) .
- Stop-Word Removal: stop-words are words that don't have any substantial meaning or any word which doesn't have any effect on meaning in terms of finding the text classification. These words must be removed from the dataset, such as conjunctions (so بالتالي, but لكن, for الى/عن/على), pronouns like (as we نحن, it هو/هي and you انت,), Prepositions like (at في/على, of عن, until حتى) [19,20].
- Stemming: The last step of pre-processing which involves removing all terms' prefixes, suffixes and infixes to extract the base root of such terms with the use an actual dictionary for terms. In our case the essays were written in Arabic , where the Arabic language has 11,347 roots. For example, stemming " اجتماعي ", "مجموعة ", " جمعة ", " جماعي ", " جامعة ", "تجمع ", "اجتماع ", "اجتماعية ", "مُجمع ", " جُموع ", "استجمع ", "جماعة ", " مجاميع ", "اجمع ", all these words have the same root "جمع". In this paper, we have used an Arabic light stemmer implemented in the WEKA tool is (Arabic light stemmer) [26] along with simple Arabic Dictionary.

3.3 Training Dataset

The training involves creating a model that is trained from by dataset to train a machine-learning algorithm by giving it a set of inputs (questions, model answers) and output (Grades). We trained a model to predict the grade of students' essays based on the features extracted from model answers using (tf-idf weights). Training data is in an important phase of machine learning where the more training data, the more the accuracy gradually increases. Performance continues to increase until the dataset is ready for testing process. In this paper, We have used three machine-learning algorithms named Decision Tree (j48), Naive Bayes, and Meta to build a model for dataset training. Truly, during the experimental evaluation of the proposed dataset, we have tried many other algorithms, and the reason why we chose such algorithms is the high accuracy of such models.

3.4 Dataset Validation and Evaluation

In order to validate the suitability of the proposed AEGD, in this paper, we have used the percentage-split technique which includes

Journal of Theoretical and Applied Information Technology
31st March 2021. Vol.99. No 6

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

partitioning the data into parts ,The dataset was divided randomly for two parts, the first part was for the training process and the second part was for the testing process. 80% of the dataset were used in the training process while 20% were used for the testing process. In this paper, we have used the precision, recall, F-measure, and ROC curve as metrics for evaluation the machine learning models on our proposed AEGD dataset.

4. EXPERIMENTAL RESULT AND DISCUSSION

The established AEGD dataset was evaluated by using three well-known machine learning algorithms: Naïve Bayes (NB), decision tree (J48), and meta classifier. We have conducted a set of experiments using the WEKA tool. The WEKA is open-source software written in Java, which compromises a set of machine learning algorithms.

The first experiment involved evaluating and testing the established AEGD using the NB classifier. The obtained results revealed that the accuracy rate for the essay questions that correctly graded is 79.0483% compared to 20.9517% incorrectly graded essays. Table 3 summarizes the metrics used to evaluate the performance of NB on established AEGD. The ROC area for NB as illustrated in Figure 3 is above the threshold. It means that NB performed well regarding predicting the grades of essays.

The second experiment involved evaluating and testing the established AEGD using the J48 classifier. The obtained results revealed that the accuracy rate for the essay questions that correctly graded is 81.2855% compared to 18.7145% incorrectly graded essays. Table 4 summarizes the metrics used to evaluate the performance of the J48 classifier on established AEGD.

The ROC area for the J48 classifie as illustrated in Figure 4 is above the threshold. It means that the J48 classifier performed well regarding predicting the grades of essays.

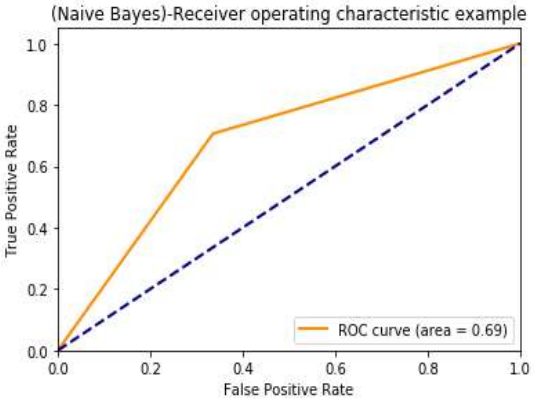

Figure 3 : The Results of the ROC Area for NB.

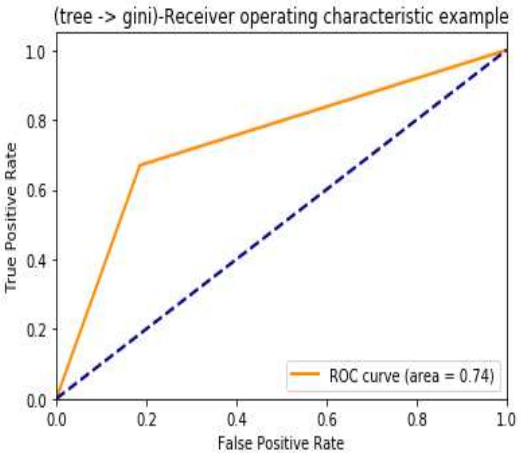

Figure 4: The Results of the ROC Area for J48.

Table 3: Evaluation Metrics Results for Naive Bayes.

	Evaluation Metrics				
	Precision	**Recall**	**F-Measure**	**ROC Area**	**Class**
NB	0.869	0.785	0.825	0.909	5
	0.795	0.810	0.802	0.882	2
	0.922	1.000	0.959	0.984	0

Table 4: Evaluation metrics results for Decision tree.

	Evaluation Metrics				
	Precision	**Recall**	**F-Measure**	**ROC Area**	**Class**
Meta	0.754	0.764	0.759	0.860	5
	0.747	0.690	0.718	0.746	2
	0.935	1.000	0.967	0.983	0

Table 5: Evaluation metrics results for Meta classifier

	Evaluation Metrics				
	Precision	**Recall**	**F-Measure**	**ROC Area**	**Class**
J48	0.849	0.755	0.799	0.895	5
	0.759	0.634	0.691	0.802	2
	0.769	1.000	0.869	0.930	0

The third experiment involved evaluating and testing the established AEGD using the Meta classifier. The obtained results revealed that the accuracy rate for the essay questions that correctly graded is 86.1151% compared to 13.8849% incorrectly graded essays. Table 5 summarizes the metrics used to evaluate the performance of the Meta classifier on established AEGD. The ROC area for Meta classifier as illustrated in Figure 5 is above the threshold. It means that the Meta classifier performed well regarding predicting the grades of essays.

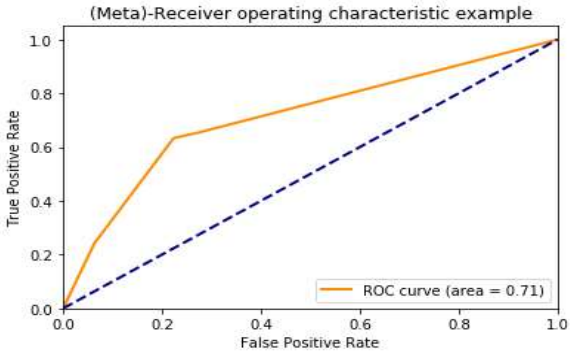

Figure 5: The Results of the ROC Area for meta.

It can be concluded from Figure 6, that the Meta classifier achieves higher accuracy in predicting the grades of essays with an 86% accuracy rate compared to the other two classifiers. Besides, the Meta classifier achieved the lowest Mean Absolute Errors (MEA) on the established AEGD as shown in figure 7.

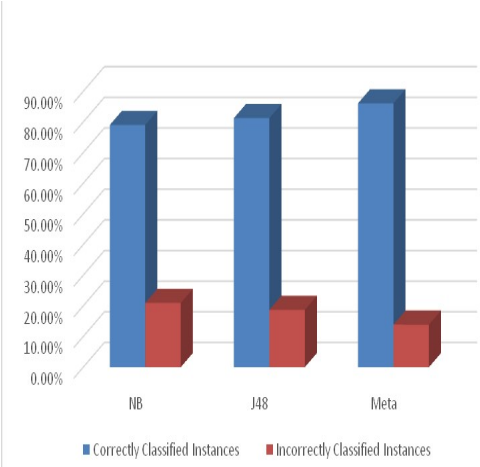

Figure 6:Accuracy of ML Algorithms based on AEGD

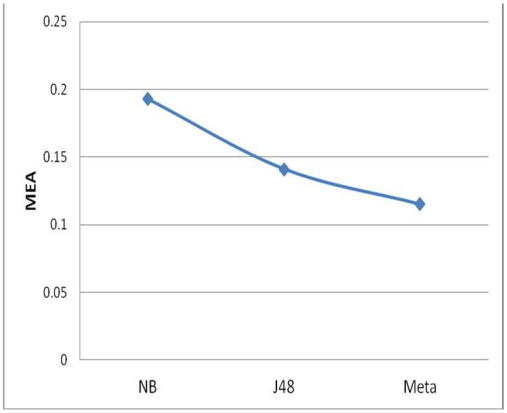

Figure 7: Mean Absolute Errors

5. CONCLUSIONS

We have introduced a new AEGD to be used to automate Arabic AEG. The established AEGD was tested and evaluated using three machine learning algorithms and the result was encouraging. The collected AEGD has taken from the teacher's guide of Islamic, History, Geography, Biology, Computer, Geology, Chemistry, and Physics. These subjects are taught in Jordanian schools from 9th to 11th grades. The collected data were analyzed and structured to ensure that the dataset is suitable for machine learning. Moreover, The dataset included 1003 questions and 3009 answers and contain a high and diverse number of Arabic unique words.

As future work, we intend to validate the dataset using deep learning approaches. In addition, we also intend to include other topics to enrich the Arabic vocabulary of the dataset.

ACKNOWLEDGMENT

The authors are grateful to the Middle East University, Amman, Jordan for the financial support granted to cover the publication fee of this research article.

REFERENCES

[1] D. S. V. Madala, A. Gangal, S. Krishna, A. Goyal, and A. Sureka, "An empirical analysis of machine learning models for automated essay grading," *PeerJ*, vol. 6, 2018, doi: 10.7287/peerj.preprints.3518v1.

[2] M. A. Hussein, H. Hassan, and M. Nassef, "Automated language essay scoring systems: A literature review," *PeerJ Computer Science*, vol. 2019, no. 8, pp. 1–16, 2019, doi: 10.7717/peerj-cs.208.

[3] V. V. Ramalingam, A. Pandian, P. Chetry, and H. Nigam, "Automated Essay Grading using Machine Learning Algorithm," *Journal of Physics: Conference Series*, vol. 1000, no. 1, 2018, doi: 10.1088/1742-6596/1000/1/012030.

[4] E. B. Page, "Project Essay Grade: PEG.," 2003.

[5] T. K. Landauer, "Automatic essay assessment," *Assessment in education: Principles, policy & practice*, vol. 10, no. 3, pp. 295–308, 2003.

[6] J. Burstein, K. Kukich, S. Wolff, C. Lu, and M. Chodorow, "Enriching Automated Essay Scoring Using Discourse Marking.," 2001.

[7] V. Santos, M. Verspoor, and J. Nerbonne, "Identifying important factors in essay grading using machine learning," *Language Testing and Evaluation Series (International Experiences in Language Testing and Assessment)*, vol. 28, no. January 2013, pp. 295–309, 2012.

[8] M. Mahana, M. Johns, and A. Apte, "Automated essay grading using machine learning," *Mach. Learn. Session, Stanford University*, 2012.

[9] M. Uto and M. Okano, "Robust neural automated essay scoring using item response theory," in *International Conference on Artificial Intelligence in Education*, 2020, pp. 549–561.

[10] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, 2015.

[11] T. O. Ayodele, "Types of machine learning algorithms," *New advances in machine learning*, pp. 19–48, 2010.

[12] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text classification algorithms: A survey," *Information*, vol. 10, no. 4, p. 150, 2019.

[13] S. Sharma and A. Goyal, "Automated Essay Grading: An Empirical Analysis of Ensemble Learning Techniques," in *Computational Methods and Data Engineering*, Springer, 2021, pp. 343–362.

[14] X. Deng, Y. Li, J. Weng, and J. Zhang, "Feature selection for text classification: A review," *Multimedia Tools and Applications*, vol. 78, no. 3, pp. 3797–3816, 2019.

[15] M. N. M. Ranjan, Y. R. Ghorpade, G. R. Kanthale, A. R. Ghorpade, and A. S. Dubey, "Document classification using lstm neural network," *Journal of Data Mining and Management*, vol. 2, no. 2, pp. 1–9, 2017.

[16] G. Wang, N. Alamas, and M. Anggraeni, "The use of internet of things and big data to improve customer data in insurance

company," *International Journal of Emerging Trends in Engineering Research*, vol. 7, no. 12, pp. 756–761, 2019, doi: 10.30534/ijeter/2019/047122019.

[17] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, "Supervised machine learning: A review of classification techniques," *Emerging artificial intelligence applications in computer engineering*, vol. 160, pp. 3–24, 2007.

[18] A. Shehab, M. Faroun, and M. Rashad, "An automatic Arabic essay grading system based on text similarity Algorithms," *Int. J. Adv. Comput. Sci. Appl.(IJACSA)*, vol. 9, no. 3, pp. 263–268, 2018.

[19] S. A. Al Awaida, B. Al-Shargabi, and T. Al-Rousan, "AUTOMATED ARABIC ESSAY GRADING SYSTEM BASED ON FScore AND ARABIC WORDNET," *Jordanian Journal of Computers and Information Technology (JJCIT)*, vol. 5, no. 03, 2019.

[20] W. H. Gomaa and A. A. Fahmy, "Automatic scoring for answers to Arabic test questions," *Computer Speech and Language*, vol. 28, no. 4, pp. 833–857, 2014, doi: 10.1016/j.csl.2013.10.005.

[21] R. Mezher and N. Omar, "A Hybrid Method of Syntactic Feature and Latent Semantic Analysis for Automatic Arabic Essay Scoring," *Journal of Applied Sciences*, vol. 16, no. 5, pp. 209–215, 2016, doi: 10.3923/jas.2016.209.215.

[22] M. M. Refaat, a a Ewees, and M. M. Eisa, "Automated Assessment of Students ' Arabic Free-Text Answers," *International Journal of Intelligent Computing And Information Science*, vol. 12, no. 1, pp. 213–222, 2012.

[23] A. M. Azmi, M. F. Al-Jouie, and M. Hussain, "AAEE – Automated evaluation of students' essays in Arabic language," *Information Processing and Management*, vol. 56, no. 5, pp. 1736–1752, 2019, doi: 10.1016/j.ipm.2019.05.008.

[24] B. Al-Shargabi, F. Olayah, and W. A. L. Romimah, "An experimental study for the effect of stop words elimination for Arabic text classification algorithms," in *International Journal of Information Technology and Web Engineering*, 2011, doi: 10.4018/jitwe.2011040106.

[25] B. Al-Shargabi, W. Al-Romimah, and F. Olayah, "A comparative study for Arabic text classification algorithms based on stop words elimination," in *ACM International Conference Proceeding Series*, 2011, doi: 10.1145/1980822.1980833.

[26] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

.