Automated Essay Scoring using Ontology Generator and Natural Language Processing with Question Generator based on Blooms Taxonomy's Cognitive Level

Jennifer O. Contreras, Shadi M. S. Hilles, Zainab Binti Abubaker

Abstract: Essay writing examination is commonly used learning activity in all levels of education and disciplines. It is advantageous in evaluating the student's learning outcomes because it gives them the chance to exhibit their knowledge and skills freely. For these reasons, a lot of researchers turned their interest in Automated essay scoring (AES) is one of the most remarkable innovations in text mining using Natural Language Processing and Machine learning algorithms. The purpose of this study is to develop an automated essay scoring that uses ontology and Natural Language Processing. Different learning algorithms showed agreeing prediction outcomes but still regression algorithm with the proper features incorporated with it may produce more accurate essay score. This study aims to increase the accuracy, reliability and validity of the AES by implementing the Gradient ridge regression with the domain ontology and other features. Linear regression, linear lasso regression and ridge regression were also used in conjunction with the different features that was extracted. The different features extracted are the domain concepts, average word length, orthography (spelling mistakes), grammar and sentiment score. The first dataset used is the ASAP dataset from Kaggle website is used to train and test different machine learning algorithms that is consist of linear regression, linear lasso regression, ridge regression and gradient boosting regression together with the different features identified. The second dataset used is the one extracted from the student's essay exam in Human Computer Interaction course.

The results show that the Gradient Boosting Regression has the highest variance and kappa scores. However, we can tell that there are similarities when it comes to performances for Linear, Ridge and Lasso regressions due to the dataset used which is ASAP. Furthermore, the results were evaluated using Cohen Weighted Kappa (CWA) score and compared the agreement between the human raters. The CWA result is 0.659 that can be interpreted as Strong level of agreement between the Human Grader and the automated essay score. Therefore, the proposed AES has 64-81% reliability level.

Keywords : automated essay scoring, linear regression, gradient boosting regression, linear lassor regression, ridge regression, gradient boosting regression, bloom's taxonomy

Revised Manuscript Received on October 05, 2019
***** Correspondence Author

Jennifer Contreras*, Computer Science, Far Eastern University Institute of Technology, Manila, Philippines. Email: jen.almarasy@gmail.com

Shadi Hilles, Faculty of Computer and Information Technology, Al-Madinah Internationa University, Kuala Lumpur. Malaysia Email: dr.shadi@bmediu.edu.my

Zainab Binti Abubakar, Faculty of Computer and Information Technology, Al-Madinah Internationa University, Kuala Lumpur. Malaysia. Email: Zainab.abubakar@mediu.edu.my

I. INTRODUCTION

In a world of continuous development and discovery of new technologies for various areas and fields to lessen the complexity of difficult tasks, education is one of the major considerations especially the assessment of student's performance. The integration of technology in learning is considered as one of the most useful and innovative techniques that enable students and teachers to collaborate effectively. Automating students' activities increases accessibility and flexibility in educational sector that enables the students and teachers to be more productive and effective.

One of the commonly used assessment to gauge the students learning is essay writing that can be applied in all levels of education and disciplines [1]. It is advantageous in evaluating the student's learning outcomes [2] because it gives the them the chance to exhibit their knowledge and skills [3]. Needless to say, it is an essential task for every language instructor to correctly score and evaluate learners' performance and proficiency. Furthermore, it will test the critical thinking skills of students if they can discuss the significance and connection between concepts that the teacher required including the vocabularies and how it fits into a larger context.

Evaluating essay depends on the objective of the exam, it could be (a) mechanics that evaluates the grammar and spelling requirements, (b) structures refers to the logical representation and construction of ideas, (c) content relates to a specific subject that an essay must contain and realize different criterion for the content, for instance, an essay may be related to some area of Artificial Intelligence in Computer Science and, the (d) style of an essay may be required to have an expository, narrative or argumentative [4].

Therefore, the score given for the essay is based on the human grader's judgment. On the other hand, the analytical evaluation identifies the characteristics of the essay content and evaluated in different ways to get the final score. To generate quicker evaluation results, the researcher used a holistic approach that makes less costly even inclined to be subjective. However, the analytical evaluation [5] is also more reliable but not cost efficient that can still be considered to use.

Retrieval Number: A9974109119/2019©BEIESP
DOI: 10.35940/ijeat.A9974.109119

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Even though the use of essay questions is beneficial to student's learning and assessment [6] the difficulties in the part of the assessors or the teachers were identified. The checking of the essay exams manually consumes a large amount of teacher's precious time [7] as the grading of essay type exam consumes a lot of time and a tedious task particularly in a huge population of students. Moreover, the discernment of the subjectivity of the scoring procedure must also be given into consideration because it may result in different score outcomes. Human graders are unique since they can have different attributes such as age, training, mood, social, backgrounds and reaction to the essay style that may influence the way they assign scores. A teacher which is considered as the human grader may be influenced by their personal know how about their students.

Automated Essay Scoring (AES) is a tool that provides objective evaluation and refrains from being subjective to save time and effort of the human graders. The main goal of AES is to automatically predict the score of students thru several features like the grammars, spelling, length of the essay etc. An AES software was introduced based from different machine learning algorithms to emulate essay scoring [8] that is deemed as its core component. According to Balfour [9], V. Paruchuri of Edx said that ES can give students feedback, correct scores and iterative on the students' essay before the response is constructed. Furthermore, AES is a student learning tool that based on the students score and provides immediate feedback which is more critical in evaluating the students' performance and how the essay can be graded based on the constructed response.

Based on the results of the literature review, it is identified that different frameworks, models or algorithms are used to deal with the automated scoring of essay exams accurately. The contribution of this study is enumerated below:

A. Proposed a machine learning framework in predicting stud on predicting student's essay score. AES is used in evaluating the essay exam of students as feature of societies and diversified composition. One of the results of this study contribute to an understanding of different machine learning models such as LASSO regression, ridge regression and gradient boosting regression to be used in training and testing AES model to perform better.

B. Improved the reliability and accuracy of assessing the essay of the students based on the AES based on Ontology and Natural Language Processing has been improved. This research study addresses different key problems in developing automated essay scoring from a unified point of view and developed a theoretical understanding of the different fundamental issues and experimental paradigm that builds this view. The theoretical work in this area contributed to developing better understanding of the different AES software introduced by different researchers with their own techniques and models. The implementation of ontology and natural language processing are the highlights in the main contribution.

C. Evaluated different machine learning models in generation essay exam question automatically based on the cognitive level of Bloom's Taxonomy. Different classifiers are tested such as Naïve Bayes, Support Vector Machine, Decision trees in classifying essay exam questions based on the cognitive level of Bloom's taxonomy.

D. Implemented the generation of essay questions for essay exam based on the cognitive level of Bloom's Taxonomy. Can a question generator be useful for the essay exam classification based on Bloom's taxonomy cognitive level for teachers? Is there a significant relationship between an essay exam question based on cognitive level of Bloom's taxonomy and the student's essay score?

This paper is structured as follows. Section 2 the review of the literature where the overview of important concepts in this study is presented. discusses and depicts the methodology of the study that starts with Data Collection, Preprocessing, Feature Extraction, Training, Testing, and Validation. 4) the fourth section is composed of evaluation and discussion about the experiments conducted and the last section 5) the last section carries the conclusion and the recommendation.

II. RELATED WORKS

This section povides the explanation of different concept domains that inspires the development of this dissertation research, then an overview showing the advantages and disadvantage features of existing AES. Moreover, the models for creating AES is examined to address different issues in developing the assessment and creating the questions for the essay exam. This chapter also contains a summary of the different algorithms applied in the study. The synthesis of the theoretical framework is also present to completely comprehend the clear picture of the research to be completed as well as the definition of the concepts used to understand the study.

A. Automated Essay Scoring

AES is an application that evaluates and predict scores [written in a text format. The notion of AES [7][8] emerged in 1966 and enhanced by Ellis Batten Page [10] who is widely known as the father of automated essay scoring. Page developed an AES called Project Essay Grader (PEG) to address the huge demand placed on teachers in evaluating essay exams. Page began with a set of student graded essay exams then experimented with a different textual feature and implement multiple linear regression. He also identified the best-weighted features combination that effectively predicts the student's essay score. Furthermore, other essay sets are graded using the same features. The process of evaluation is based on concept or "proxes". These "proxes" are variables of interest within essay such: count of proposition or complexity of special words. Hence, the overall performance score of PEF is the correlation of 0.87 with human graders that simply means it has a good correlation result.

Table- I: Survey of AES Techniques

Author	Name	Scoring Model	Scoring Categories
Ellis Page	Project Essay Grade (PEG) 1966	Regression model: used proxes and trins	Writing quality, Organization support, sentence structure, word choice, mechanics
Peter Foltz and Thomas Landauer	Intelligent Essay Assessor 1989	Latent Semantic Analysis (LSA)	Identify which numerous calibration documents are most like the essay, most similar calibration documents are assigned to the AES score
Vantage Learning's	IntelliMetric (1999)	Probabilistic model	Common domains of writing: Emphasis, Composition, Development, Mechanics, Grammar, Voice, and Language Use.
Jill Burstein (2001)	E-rater	Natural Language Processing	Content analysis based on vocabulary measures lexical complexity /diction, the proportion of grammar errors, the proportion of usage errors, the proportion of mechanics errors, proportion of style comments, organization and development scores, features rewarding idiomatic phraseology
Lawrence Rudner (1996)	BETSY	Naive Bayes	Multinomial or Bernoulli Naive Bayes models, Optional Porter stemming, Re-entrant training, Popular database format, Output on screen and into CSV files, Training file trimming, Infrequent term purge, Diagnostic information, up to 5 score categories and Free for non-commercial use.
Kaggle	Automated Student Assessment Prize (ASAP)	Random Forest	Scored by annotator: annotations are scores for different attributes of the essays, such as content, word choice, organization, sentence fluency, etc.

[a] Sample of a Table footnote. (*Table footnote*)

Table shows the survey of the common AES software, model and the features they considered in scoring the essay. Until this day, one of the major challenges in the AES is the deficiency of AES system that will impart insights into their grading methodology the mere reason why most of them are created by a commercial entity [10].

A. Ontology

Ontology simply the concept map of domain knowledge. It is originally borrowed from Philosophy which means “Theory of existence” that obtained considerable recognition in Computer Science and Information System since it targets one of the core difficulties of using computer for human purpose which is achieving interoperability between multiple representations of reality such as data or business models, and within those representations and reality, namely human users and their perception of reality. According to Tom Gruber, ontology is a “specification of a conceptualization” [12] that gives us the relations that exist between annotations, helps us understand each annotated token in a larger context, and helps us understand what information is missing and what else we need to look for. Therefore, we can say that the use of ontology in AES can enhance the performance of machine learning techniques [13].

B. Cognitive Level of Blooms Taxonomy

Blooms Taxonomy was introduced by Benjamin Bloom with his team members, Max Englehart, Edward Furst, Walter Hill and David Krathwohl [14] in 1956. It is a product of a sequence of conferences in higher education that focuses on enhancing the communication and organisation in doing assessments in education sector. The taxonomy is represented in hierarchical form starting from the lowest level elevated in a more advanced higher-order learning objectives beyond the basics of remembering and understanding. Introduced as a type of classifications of learning outcomes and objectives, BT cognitive level was used for everything from framing digital tasks and evaluating app to writing questions and assessments. It is a hierarchical ordering of cognitive skills that can help teachers teach and students to learn. Several academic governing bodies like accreditation are checking the correctness of the classification of exam questions since some of the teachers
who’s creating the exam questions is not familiar with the proper adaptation of the cognitive classification of BT of learning.

The original taxonomy includes the following categories: “knowledge, comprehension, application, analysis, synthesis, and evaluation”. Meanwhile, the initial effort of the participants in the conference and others have develop a number of revisions on the taxonomy [15].

In 2001, Anderson and Krathwohl published the revised taxonomy that is composed of (1) remembering, (2) understanding, (3) applying, (4) analyzing, (5) evaluating and (5) creating as presented in Figure 1.3.

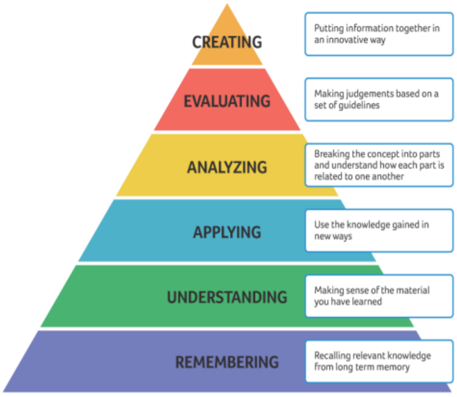
CREATING
Putting information together in an innovative way
EVALUATING
Making judgements based on a set of guidelines
ANALYZING
Breaking the concept into parts and understand how each part is related to one another
APPLYING
Use the knowledge gained in new ways
UNDERSTANDING
Making sense of the material you have learned
REMEMBERING
Recalling relevant knowledge from long term memory

Figure 1.0 Bloom's Taxonomy-Guest Blog: Wi-fi Certification Training Essentials From Ekahu Wireless Design, by K. Parsons, 2016,https://www.ekahau.com/blog/2016/12/13/guest-blog-wi-fi-certification-trainingessentials/://mramusicplace.net/2017/01/12/music-teaching-and-blooms-revised-taxonomy/. Copyright 2016

Fig 1.0 is an illustration of the Revised Six Cognitive Level of Bloom's Taxonomy by Caryn McKindraw (2019) starting from the lowest level which is (1) ***remembering*** that simply means recalling the significant knowledge from long term memory, (2) ***understanding*** is a level of learning making sense of the material you have learned, (3) ***applying*** which is basically using the knowledge gained in new ways, (4) ***analyzing*** is breaking the concept into parts and understand how each part is related to one another, (5) ***evaluating*** means making judgements based on a set of guidelines, and is (6) ***creating*** which means putting information together in a more innovative way.

A. Machine Learning Algorithms

- **Support Vector Machine** (SVM) is a supervised classification algorithm that was officially introduced by Boser, Guyon, and Vapnik (1992) [16] during the Fifth Annual Association for Computing Machinery Workshop on Computational Learning Theory. Nowadays, it is considered as a powerful classification and regression tools [17]. It has been deployed in a wide range of real-world problems such as text categorization, handwritten digit recognition, tone recognition, image classification, object detection and data classification [18].

The SVM illustration is show below:

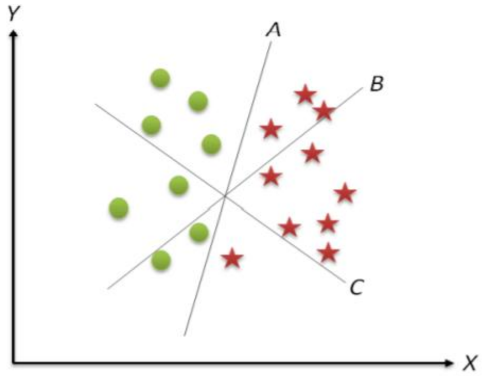

Figure 2.0 SVM Machine Learning model

SVM is commonly applied in a problem that involves classification such as pattern recognition, image classification, text classification and AES [19]. Just like other machine learning algorithms, SVM also require pre-selected features such as word length, word level, spelling errors, sentence length, and sentence level [20] that can be done using NLP. That is why we can say that the tasks in NLP usually represents instance by very high dimensional but very sparse feature vectors, which leads to positive and negative examples being distributed into two different areas of the feature space. In fact, this is useful to the SVM to look for the classification hyperplane in feature space and for the generalization capability of the classifier. It is the mere reason why SVM can attain better results in different NLP applications.

- **K-Nearest Neighbor (KNN)**

KNN is called lazy algorithm [21] since it is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). It has been used in statistical estimation and pattern recognition already in the beginning of 1970s as non-parametric technique. KNN stores all available cases and classifies new cases based on similarity measures like the distance functions. Just like other machine learning algorithms, KNN can also be implemented for classification and regression predictive problems. However, it is more applicable in classification problems in the industry because it easier to interpret the output, fast calculation time and its predictive power.

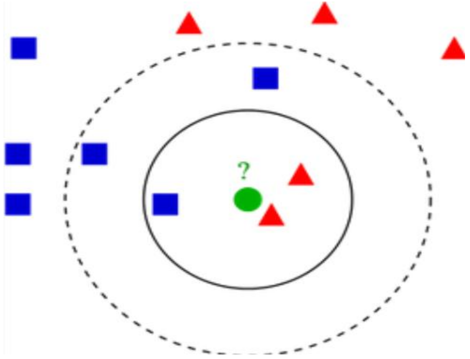

Figure 3.0 KNN Machine Learning model

Figure illustrates the KNN classification, given the test sample is the green circle should be classified either to the first class of blue squares or to the second class of red triangles. If k = 3 (solid line circle) it is assigned to the second class because there are 2 triangles and only 1 blue square inside the inner circle. On the other hand, if K=5 (dashed line circle) it is assigned to the first class since there are 3 blues squares and 2 red triangles inside the outer circle.

- **Naïve Bayes (NB)**

Classification and prediction are two the most important aspects of Machine Learning and NB is a simple but surprisingly powerful algorithm for predictive modelling.

It is a probabilistic classifier [22] which is created based on Bayes theorem with naive assumptions regarding independence

between every pair of features. Assume a variable C denotes the class of an observation O. The class of the observation O can be predicted using the Bayes rule, we need to calculate the highest posterior probability of [23].

The Naïve Bayes theorem:

$$P(C|O) = \frac{P(C)P(C|O)}{P(O)} \qquad (2.1)$$

In the NB classifier, using the assumption that features O1, O2, ..., On are conditionally independent of each other given the class, we [23].

$$P(C|O) = \frac{P(C)\prod_{i+1}^{n} P(O_i|C)}{P(O)} \qquad (2.2)$$

III. AES FRAMEWORK

In this chapter, the dataset used in the current study, the development of the model and architecture and the evaluation measures used for model validation are discussed in detail. Various problems and solutions are mentioned in the previous chapters that motivates the development of this research study. The methodology is based on the overview and the findings of the different AES as stated on the second chapter becomes the basis of the development from the data gathering to the evaluation techniques of the AES.

A. Overview of the System Framework

The study is divided into three main phases and each phase generates an output that is important as an input for the succeeding phases. The overview of the research framework for design and development of the proposed automated essay scoring using ontology and NLP with the automatic generation of essay exam based on the cognitive level of Blooms Taxonomy. The overview of the research framework for design and development of the proposed study is illustrated in figure below.

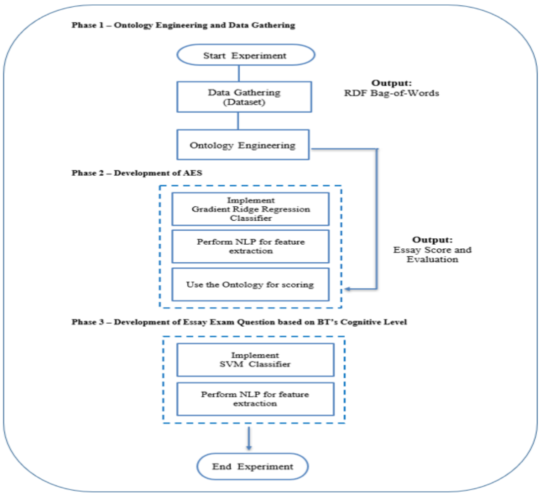
Phase 1 – Ontology Engineering and Data Gathering
Start Experiment
Data Gathering (Dataset)
Output:
RDF Bag-of-Words
Ontology Engineering
Phase 2 – Development of AES
Implement Gradient Ridge Regression Classifier
Perform NLP for feature extraction
Use the Ontology for scoring
Output:
Essay Score and Evaluation
Phase 3 – Development of Essay Exam Question based on BT's Cognitive Level
Implement SVM Classifier
Perform NLP for feature extraction
End Experiment

Figure 4.0 System Framework

Fig illustrates the system framework of the proposed AES that begins with phase one (1), the ontology engineering. The creation of ontology is important since the scoring of the vocabularies for the essay will based from the similarity of the BOW from the essay of the students. Moreover, the actual development of the AES in phase two (2) starts with the preprocessing phase that cleans and prepare the data for feature extraction and selection that will be used by different classification algorithms to determine which is the best regression algorithm to implement. In this study, various algorithms for classification are tested to identify the excellent model to be implemented in the system. Scores were predicted thru different features extracted and the domain concepts from the ontology.

Finally, phase three (3) which the last phase in the experiment focuses on the development of the generation of essay exam question based on BT's cognitive level. This will be helpful to the teachers to familiarize the creation of essay exam base on BT's cognitive level. After testing the application, different essay exam questions were collected to check its accuracy.

B. Ontology Engineering and Data Gathering

Phase one (1) has two main research activities which concerns on gathering or selecting of datasets to be used in the experiment and the engineering of ontology. To evaluate the students, essay score different datasets were collected including the ASAP dataset from Kaggle website since the "training set needs to be large enough to allow the tool to generalize to the population of expected responses it will score" (Foltz et al., 2012). On the other hand, this study also considered the use of Kaggle dataset that is helpful for training the model. Moreover, a corpus is also used to extract the BOW which is a set of "naturally-occurring language text" that is selected to describe a different state a language. It is a huge amount of linguistic evidence that is comprised of the implemented language.

C. Automated Student Assessment Prize (ASAP)

ASAP dataset is available at Kaggle website that is commonly used by researchers who are working with AES. It was used for the Kaggle competition to write an AES based on their own criteria that includes using the datasets they have provided that is composed of eight (8) essay datasets. Every dataset was generated from a single prompt and the essays have an average length of 150 to 550 words per answered essay exams. Moreover, the gathered answers were written by students from grade 7 to grade 10 and all are checked manually and double checked. Each of the eight data sets has its own unique characteristics including the types of essay: persuasive, narrative, expository and source dependent responses.

Table- II: Details about ASAP Dataset

PROMPT ID	Essays	Score Range
1	1783	2-12
2	1800	1-6
3	1726	0-3

4	1772	0-3
5	1805	0-4
6	1800	0-4
7	1569	0-30
8	723	0-60

The eight (8) datasets are the eight (8) prompts and12976 essays, refer to the table below for more details about the ASAP dataset.

D. Essay scores from an essay exam in HCI and Software Engineering courses

Real essay exam was collected from the students with their corresponding scores from different courses. The essays were selected based on the score of the students. The score must have a grade of 0 to 5 based on the TOEFL examination score.

E. Evaluating the Performance using Kappa Coefficient (Cohen's kappa correlation)

Cohen's Kappa is a statistical measure created by Jacob Cohen in 1960 to measure the reliability accurately between two raters making decisions about how a unit of analysis should be categorized. This statistical measure between raters for categorical type of data. It measures not only the % of agreement between two raters, it also calculates the degree to which agreement can be attributed to chance. This measure was applied by Jill Burstein in their experiments where in an estimated of agreement was to be provided for machine evaluations and human graders for text essay.

The mathematical formula for Cohen's Kappa:

$$K = \frac{Pr(a) - Pr(e)}{N - Pr(e)} \qquad (1)$$

Where:

- *Pr(a)* refers to the simple agreement among raters
- *Pr(e)* refers to the probability that agreement is attributable to chance
- *N* is the total number of rated items, also called "cases"

Kappa statistic is commonly applied if we need to test the reliability of the interrater which important because it signifies the degree to which the data that was collected for this research are appropriate to represent the variables measures. Moreover, interrater reliability is a kind of measures of the degree in which the one who's collecting the data which is the rater designates the same score to the same variables. Though there are different methods in measuring reliability among interrater, conventionally, it is measured in percentage agreement that computes the number of agreement scores divided by the total number of scores. Like the other correlation statistics, the Cohen kappa's result is ranging from −1 to +1.

Table- II: Cohen's Kappa Interpretation

Kappa Value	Level of Agreement	% of reliable data
0 - .20	Minimal	0 – 4%
.21 - .39	Weak	4 – 15%
.40 - .59	Moderate	35 – 63%
.60 - .79	Strong	64 – 81%
.90 – 1.0	Almost Perfect	82 – 100%

F. Precision

This measure is useful here for comparison as we have only two cases of agreement i.e. the human and machine grading agrees or does not agree. In this study, a precision is used to evaluate the essay results that is represented below:

$$P = \frac{No.of\ cases\ of\ agreement}{No.of\ evaluations\ by\ Human\ +\ No.of\ evaluations\ of\ Machine}$$

IV. EXPERIMENTAL RESULTS FOR AES

Before you begin to format your paper, first write and save the content as a separate text file. Complete all content and organizational editing before formatting. Please note sections A-D below for more information on proofreading, spelling and grammar.

Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

A. Experiement#1. Results and Analysis for AES based on Ontology and Natural Language Processing using ASAP dataset

After the features have been extracted, different data prior to the implementation of the different classification models: Linear Regression, LASSO Regression, Ridge Regression and Gradient Boosting Regressing for training and testing. These labels were implemented for training the classifiers. The accuracy table below shows the result of the accuracy and the validation (Kappa) based on the results generated by the classifiers.

Table- III: Result of Training a different regression modelusing 10 numerical, POS and orthographic features

Training Model	Cohen's Kappa Score	Mean squared error	Variance score
Linear Regression	0.32	0.69	0.66
LASSO Regression	0.34	0.69	0.66
Ridge Regression	0.32	0.69	0.66
Gradient Boosting Regression	0.40	0.6114	0.70

Table III shows the statistical results of the different regression models using 10 numerical, POS and orthographic features. Based on the given results, we can say that Gradient Boosting Regression has the highest variance and kappa scores. However, we can tell that there are similarities when it comes to performances for Linear, Ridge and Lasso regressions.

B. Experiment#2: Evaluation of the Proposed Automated AES using the Collected Essay with score from Students

Experiment two is the testing and implementation of the algorithm that was identified in the first experiment. Once the creation is done by the teacher or ontology engineer. The essay exam results were collected from the students of Computer Science in FEU Institute of Technology in Human Computer Interaction course based non “Human Interaction” topic.

To start the experiment, the teacher was asked to use the OntoGen software to create the domain ontology by uploading different essays from the student and applied the unsupervised and supervised learning and the result is shown below:

Table- IV: Similarity of concepts

Training Model	Similarities
Supervised	47.16%
Unsupervised	34.65%

Table IV shows that the average similarity for supervised learning is 47.16% out of the 18 essays as compared to the unsupervised learning which is 34.65%, therefore, we can say that supervised learning is better than unsupervised learning in terms linear regression.

On the other hand, the domain ontology is visualized below:

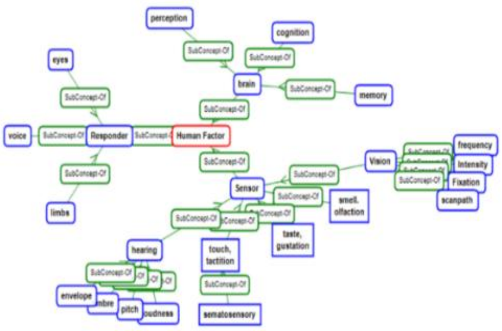

Figure 5.0 Domain ontology extracted from the essay exams of students

Figure show the visualization of the different domain ontologies extracted for the essay documents.

The extracted features are:

Table- V: Features Extracted from the HCI Dataset

Essay	Character Count	Word Count	Sentence Count
Essay1	1995	194	11
Essay2	128	11	1
Essay3	1096	112	18
Essay4	1208	123	16
Essay5	1413	140	5
Essay6	160	15	1
Essay7	1295	117	12
Essay8	153	11	1
Essay9	228	21	2
Essay10	1103	97	6
Essay11	1202	139	8
Essay12	3279	299	16
Essay13	1338	128	7
Essay14	1585	150	4
Essay15	2221	188	8
Essay16	945	99	7
Essay17	1775	167	12
Essay18	1157	112	6
Average	**1238**	**118**	**8**

Table V shows some features that were extracted from the text such as character count, word count and sentence count that was extracted from the essay. In this study, we hypothesized that counting of words from the essay would certainly be correlated positively with the essay score. We extracted the total character, words and sentences as represented on the table 2, however we only considered “word count” against the human score since words correlates to the vocabulary.

The other features included are grammar count, spelling count, sentiment polarity score etc.

The Quadratic Weighted Kappa result is to measure the inter-reliability.

Table- VI: Cohen’s weighted kappa score

	Value	Asymp. Std. Error^a	Approx. T^b	Approx. Sig.
Measure of Agreement Kappa	**.659**	.118	.118	6.191

N of Valid Cases	375			

Table 4.7 shows the CWA result is 0.659 that can be interpreted as Strong level of agreement (as stated in Chapter 2) between the Human Grader and the automated essay score. Therefore, the proposed AES has 64-81% reliability level.

V. RESULT AND DISCUSSION

The contents of the journal are peer-reviewed and archival. The journal publishes scholarly articles of archival value as well as tutorial expositions and critical reviews of classical subjects and topics of current interest.

Authors should consider the following points:

Table- V: Result of SVM with classical feature TF-IDF

Cognitive Level	Recall	Precision	F-measure
Remembertin g	0.900	0.868	0.884
Understandin g	0.861	0.897	0.876
Applying	0.741	0.873	0.798
Analyzing	0.950	0.725	0.821
Evaluating	0.809	0.857	0.830
Creating	0.770	0.889	0.825
Average	0.839	0.851	0.839

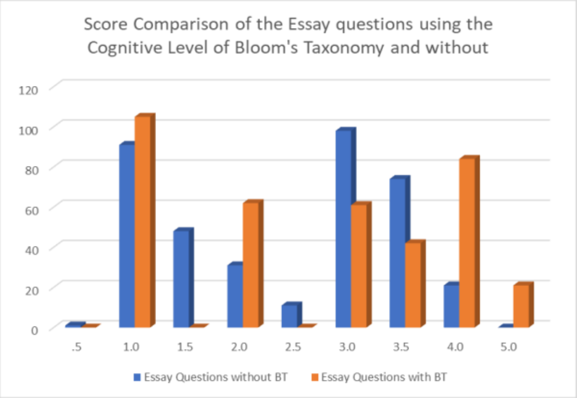

Figure 6.0 Domain The result of the score comparison of the essay questions using the cognitive level of Bloom's Taxonomy and without

Figure 5.1 illustrates the result of the comparison between the student's essay score who answered questions based with cognitive level of Bloom's taxonomy and those without. The graph shows that majority of the essay scores appears to be higher in the essay questions with Bloom's taxonomy.

VI. CONCLUSION

This research presented a simple algorithm to compute the essay score automatically based on domain ontology and NLP with the automated essay exam generator based on the cognitive level of Bloom's taxonomy. Based on the results presented in Chapter 4 and Chapter 5, it is believed that the research achieved all the objectives presented in Chapter 1 that includes: *RQ1. To study and identify the most accurate machine learning algorithm for training and testing data in an automated essay scoring and classification of the essay exam question based the cognitive level of Bloom's taxonomy*. The researcher performed two experiments to investigate the best machine learning algorithm to be used in training the data and the researcher found out that the implementation of the different classifier model differs from the features used in the experiment. For instance, during the first (1) experiment, the researcher only used one feature which is the extracted BOW from the ASAP dataset then and the Linear Regression and Linear Lasso Regression. Here, the researcher found out that the Linear Lasso Regression is the better machine learning algorithm to be used because it performs well having the Kappa score of 0.23 as compared to Linear Regression having the Kappa score of only 0.12. For the second (2) experiment, a combination of features (word count, wrong spelling count, grammatical count) and BOW using Linear Regression, Linear Lasso Regression, Ridge Regression and Gradient Boosting Regression. On this experiment, the Gradient Boosting Regression having the score of 0.37 is considered as the best machine learning algorithm to train the model though the scores of all the regression algorithms falls on the "Weak" level of agreement and the percent of data that are reliable is only form 4-15%. Furthermore, for the result of training a different regression models using 10 numerical, POS and orthographic features, the researcher found out that still the Gradient Boosting got the highest score which is 0.61 as compared to the three models having the same score of 0.69. Therefore, we can say that after the experiment implemented in this study the Gradient Boosting Regression algorithm is the best algorithm to be used to train the model.

On the other hand, to achieve the research objective, *RQ2. To identify the accuracy and reliability of the different machine algorithm models for automated essay scoring based and automated generation of exam question base on the cognitive level of Bloom's taxonomy*. This study discovered to assess the impact of using ontology in the essay scores of the students that uses features extracted from the student's essay to improve the correlations with human scores and the AES. The results of Cohen Kappa Scores shows that the result of the correlation varies depending on the features used to assess the model. Based on the evaluation done in this study, the researcher found out that Cohen's Kappa gives a significant promising result. Therefore, we can say that the overall computation of Cohen's Kappa score of the Gradient Boosting Regression algorithm which is 0.40 is interpreted with "Moderate" level of agreement and the percentage of data reliability is 35 – 63%.

Furthermore, *R3.To develop the automated essay scoring with essay exam generator based on the cognitive level of Bloom's taxonomy.* The functional web application was developed to perform the generation of essay exam question based on BT's cognitive level and the prediction of the student's essay exam. The functionality and usability were evaluated, and the result is The overall mean is 3.584667 which means that the 30 faculty members of the College of Information Technology at FEU Institute of Technology in the "Satisfactory" level of using the software for generation the essay exam question based on Bloom's Taxonomy. Therefore, we can say that software is "Satisfactory", but it needs more improvement to make it better.

Lastly, in *RQ4. To implement and evaluate the selected machine learning model for automated essay scoring and the automated generation of essay question base on the cognitive level of Bloom's Taxonomy.* The CWA result is 0.659 that can be interpreted as Strong level of agreement (as stated in Chapter 2) between the Human Grader and the automated essay score. Therefore, the proposed AES has 64-81% reliability level.

REFERENCES

1. P.Phandi, K.M. Chai, & H.T. Ng "Flexible Domain Adaptation for Automated Essay Scoring Using Correlated Linear Regression," Conference on Empirical Methods in Natural Languages Processing, pages 431-439, Lisbon, Portugal, 17-21 September 2015. Association for Computational Linguistics.
2. K. Zupanc & B. Zoran "Advances in the Field of Automated Essay Evaluation," Informatica 39 (2015) pp. 383-395.
3. S. A. Crossley, E.L. Snow, L.K. Allen & D. McNamara "Incorporating learning characteristics into automatic essay scoring models: What individual differences and linguistic features tell us about wiring quality," . Journal of Educational Data Mining, Volume 8, No. 2, 2016
4. A. Smolentsov, R. Östling , B.T. Hinnerich & E. Hoglin "Automated Essay Scoring for Swedish. Eight Workshop on Innovative Use of NLP for Building Educational Applications," pages 42-47, Atlanta, Georgia, June 13, 2013. Association for Computational Linguistics
5. T. K. Ghalib & A. Al-Hattami "Holistic versus Analytic Evaluation of EFL Writing: A Case Study,". Canadian Center of Science and Education, English Language Teaching; Vol. 8, No. 7; 2015
6. A. Groza & R. Szabo "Enacting textual entailment and ontologies for automated essay grading in chemical domain,". 16th Int. Symposium on Computational Intelligence and Informatics CINTI2015), Budapest, Hungary, 19-21
7. V. V. Ramligan, A. Panidan, P. Chetry & N. Nigam "Automated Essay Grading using Machine Learning Algorithm. National Conference on Mathematical Techniques and its Applications (NCMTA 18),". National Conference on Mathematical Techniques and its Applications (NCMTA 18). IOP Conf. Series: Journal of Physics: Conference, Series 1000 (2018) 102030. doi :10.1088/1742-6596/1000/1/012030.
8. P. J. Sijimol & M.V. Surekha "Short Answer Scoring System Using Neural Networks,". International Journal of Computer & Mathematical Sciences (IJCMS) ISSN 2347-8527 Volume 7, Issue 4 April 2018.
9. S. P. Balfour "Assessing Writing in MOOCs: Automated Essay Scoring and Calibrated Peer Review,". Research and Practice Assessment Volume Eight - Summer 2013
10. D. Ambebaker "Evaluation of essays using incremental training for Maximizing Human-Machine agreement (master's thesis)," Indian Institute of Technology, Bombay
11. K. Zupanc & B. Zoran "Advances in the Field of Automated Essay Evaluation," . Informatica 39 (2015) pp. 383-395.
12. F. Neuhas "What is Ontology?" . arXiv:1810.09171v1 [cs.AI] 22 Oct 2018
13. S. Devi & H. Mittal "Machine Learning Techniques with Ontology for Subjective Answer Evaluation" International Journal on Natural Language Computing (IJNLC) Vol. 5, No.2 2016
14. V. Wang & T. Neighmann "English Teaching and Andragogy in Transitioning Students from Secondary to Higher Education in China." 2017
15. W. N. Arlianty, B.W. Febriana, A. Diniaty & L. Fauzi'ah "Student profile in completing questions based on cognitive level of bloom's taxonomy by Anderson and Krathwoh." AIP Conference Proceedings 2026, 020063 (2018) ; https://doi.org/10.1063/1.5065023
16. D. Abduljabbar & N. Omar "Exam Questions Classification based on Bloom's Taxonomy Cognitive Level using Classifiers Combination." Journal of Theoretical and Applied Information Technology (JATIT) Volume 78 Number 3 2015
17. A. Suresh & M. Jha "Automated Essay Grading using Natural Language Processing and Support Vector Machine." IJCAT – International Journal of Computing and Technology, Volume 5, Issue 2, February 2018.
18. K.S. Duresh & B. Lekha Data Classification using Support Vector Machine. Journal of Theoretical and Applied Information Technology". Journal of Theoretical and Applied Information Technology, 2010. 12(1): p.1-7.
19. K.A. Osadi, M.G. N.A.S. Fernando & W.V. Welgama " Ensemble Classifier Based Approach for Classification of Examination Questions into Bloom's Taxonomy Cognitive Levels". International Journal of Computer Applications (0975-8887) Volume * - No. *, February 2017.
20. A Suresh & M. Brenner " A Least Absolute Shrinkage and Selection Operator (LASSO) for Non-Linear System Identification 2006
21. M. Mohamed, & N. Omar "Question Classification Based on Bloom's Taxonomy Using Enhanced TF-IDF.," International Journal on Advanced Science Engineering Information Technology.
22. A. Osman, & A. Yahya "Classification of Exam Questions Using Linguistically-Motivated Feature: A Case Study based on Bloom's Taxonomy," The Sixth International Arab Conference on Quality Assurance in Higher Educations (IACQA'2016)
23. T. Sonat & M. Musa "Learning the Naïve Bayes Classifier with Optimization Models," International Journal of Applied
24. Mathematics and Computer Science, vol. 23, no. 4, pp. 787–795. 2013

AUTHORS PROFILE

Jennifer O. Contreras is pursuing her Doctorate degree in Doctor of Philosophy in Information & Communication Technology and research is ongoing on automated essay scoring using ontology and natural language processing. She is also taking up Doctor of Philosophy in Computer Science with an ongoing research in ontology engineering using enhanced SVM and Artificial Neural Network. Shas been teaching for almost 20 years in different countries with more than 15 publications so far. Her research area includes Natural Language Processing, Data Mining, Ontology, Machine Learning and Image Processing.

Shadi M. S. Hilles is an Associate Professor, Computer Science Department and Project Manager in Research and Innovation in Al-Madinah International University, currently working on Project Industrial Revolution 4.0 Professional Education, research & Innovation", Dr. Shadi was keynote speaker, committee member and advisor committee in several international conference and chairman in Seminar in ICT field and area of IR 4.0 and smart campus and has published in ISI and Scopus indexed journals in computer vision fields and cryptography network security, also published in indexed book chapters in Springer and IGI in area of 5G wireless network, Dr. Shadi has presented papers in several conferences included IEEE and was one of track chair of IEEE conference and presented in numerous workshops in field of computer science and He is a main supervisor for PhD and Master students and was examiners for postgraduate students, under his supervision there are four PhD students have graduated in image processing field and wireless networking, six master students in image processing and in cryptography and network security, he is one of research development team and editorial manager in industrial revolution 4.0 Professional Education, Research & Innovation in MEDIU and Professional membership of Institute For Engineering Research and Publication, and also in SPIE-The international Society for Optical Engineering.

Dr. Zainab Binti Abu Bakar finished Doctor of Philosophy in Computer Science in 1999 at the University Kebangsaan, Malaysia and her master's degree in computer science in 1988 at Loyola Marymount University, Los Angeles, U.S.A. She has more than 100 research publications in Information Retrieval – Text Information Retrieval, Spoken Document Retrieval, Geographical Information Retrieval, Mathematical Information Retrieval, Genomic Information Retrieval, Multimedia Information Retrieval (Image, Video and Audio) and Natural Language Processing– Part of Speech Tagging, Topic Extraction and Summarization, Spoken Word Segmentation and Recognition, Lexicon Construction, Grammatical Analysis.