Automated Essay Grading using Machine Learning Algorithm

IOP Conf. Series: Journal of Physics: Conf. Series **1000** (2018) 012030 doi:10.1088/1742-6596/1000/1/012030

Automated Essay Grading using Machine Learning Algorithm

V. V.Ramalingam[1] , APandian[2], Prateek Chetry[3]and Himanshu Nigam[4]

Department of Computer Science and Engineering, S.R.M. Institute of Science and Technology, Chennai, India - 603203

ramabi1976@gmail

Abstract. Essays are paramount for of assessing the academic excellence along with linking the different ideas with the ability to recall but are notably time consuming when they are assessed manually. Manual grading takes significant amount of evaluator's time and hence it is an expensive process. Automated grading if proven effective will not only reduce the time for assessment but comparing it with human scores will also make the score realistic. The project aims to develop an automated essay assessment system by use of machine learning techniques by classifying a corpus of textual entities into small number of discrete categories, corresponding to possible grades. Linear regression technique will be utilized for training the model along with making the use of various other classifications and clustering techniques. We intend to train classifiers on the training set, make it go through the downloaded dataset, and then measure performance our dataset by comparing the obtained values with the dataset values. We have implemented our model using java.

1. Introduction

The impacts that computer have on our writings have been in use for 40 years now. Even the most basics of computers like processing of words, is of great help to authors in updating their writing material. The research has revealed that computers have the capacity to function as a more effective cognitive tool. Revision and feedback are important part of the writing material. Students in general require input from their teachers for mastering the art of writing. However, responding to student papers can be a burden for teachers. Particularly giving feedback to a large group of students on frequent writing assignments can be pretty hectic from the teacher's point of view. We are creating a system that can be accurate in both providing the feedback as well as grading the performance and can also be time efficient. In our system the scores would be much more detail oriented than the ratings provided by two human raters. Computerized scoring has many weak points as well. The method used by Hamp-Lyons stated the lack of man to man communication as well as the sense in which the writer rates the essays. Similarly, Page stated that the computers could not assess an essay as human raters do because computer systems are programmed and lack the intensity of human emotions and therefore it will not be able to appreciate the context. Another criticism is the construct objections. That is, computer can give importance to unimportant factors while rating or providing the score to the user i.e., focusing on traditional aspects rather than orthodox ones.

2. Problem Statement

Analyzing natural language, or free-form text used in everyday human-to-human communications, is a vast and complex problem for computer machines irrespective of the medium that has been selected, be it in the form of verbal interaction, written sample, or reading. The murkiness in the given system language and the lack of one definite output to any given communication task make grading, evaluating or scoring a difficult challenge. In general, in this particular domain implementing machine learning techniques with various features spaces, and large collection of data having different pattern can give us better outcomes.

2.1 *Disadvantages of ExistingSystem*

The existing system implements linear regression model to learn from these features and generate parameters for testing and validation.

a) Lack of autocorrelation
b) Homoscedastic (variance independent ofpoint)
c) Nomulticollinearity

3. Related Work

3.1 Automated Essay Grading in IndianContext

Automated Essay Grading (AEG) or scoring has turned into a reality now. Artificial Intelligence systems provide a lot to the educational community where graders have to face different kind of difficulties while rating student writings. Analyzing student essays in abundance within given time limit, along with feedback is a challenging task. But with changing times, the human written (not hand written) essays are easy to evaluate with the help of AEG systems. The Graduate record examination (GRE) is one of the best results of this method. The students' writings are graded by both human and automated essay grading system. Then the average istaken.

3.2 Efficient Statistical Validation for Autonomous Driving

Today's automotive industry is making a gradual shift in technology and thereby equips vehicle with intelligent driver assistance features. A modern automobile is now provided with a strong AI in order to run multiple machine learning algorithms for environmental safety and motion control. These machine learning systems must be highly sturdy with extremely small error rate to ensure safe and reliable driving. In particular, a Markov Chain Monte Carlo algorithm which is created on graphing is displayedto exactly provide the rare wrong rate with a little amount of data given, therefore decreasing the cost used in validating the method.

3.*3 Exploiting Randomness in Sketching for Efficient Hardware Implementation*

Different researches are going on that are working on efficient processing of big matrices in applications using hardware enhancement. Showing huge matrices into their lower form is a new way of working on it. This paper uses a highlyeffective hardware working of a class of forcing down methods depend on unknown values, called as Johnson Lindenstrauss (JL) transform. Basically, we show how the randomness given in the given matrix can be used to create an efficient system using machine learning technique. In particular, the transform's random matrices have two main features that provide for various results. The first feature helps us to reduce operation on different widths in calculation of the JL transform. The second feature is the property that helps to reduce width in the current machine learning techniques.

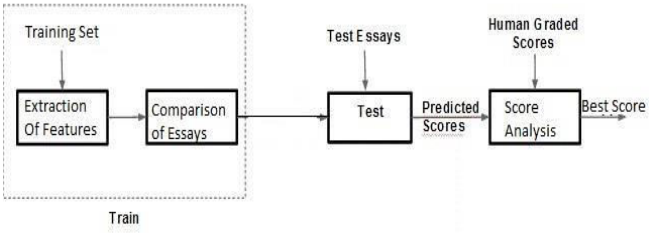
4. System Architecture

Feature extraction plays a significant role in any of the machine learning task and therefore here also we will utilize the concept of it. To build a robust yet effective essay scoring algorithm, our aim is to bank on model attributes like language fluency, grammatical and syntactic correctness, vocabulary and types of words used, essay length, domain information etc. At present, our model is making use of the

following set of features extracted from the ASCII text of the essays:

4.1 *Word count and Sentence count*
These are trivial features of any text document but it influences the scoring of the document as well. In order to extract these features, we will make use of the "text mining" library using this library, we extracted the term-document matrix so that we can provide data to our train corpuses (8900 documents). In this library a list of 276 common stop words in English language is also provided. Now, since these stop words are of not much use, we can skip them while we are calculating the word count of each document from the term document matrix. To get the sentence count, we simply break the document using '.' and thus, count the number of segmentsobtained.

Figure 1. System Architecture



4.2 POS Tag
To evaluate the quality of content in an essay is the foremost important task. Along with it an another important set of features for evaluating any piece of writing provide is the number of words in various syntactic classes like nouns, adverbs, verbs, adjectives etc. should also be evaluated . To get the counts of words in each POS (part-of-speech) class, we use NLTK library (http://nltk.org/). This library provides us the POS tag for each word in an essay, and thus, we can extract the number of nouns, adverbs, adjectives and verbs separately.

4.3 Spelling Mistakes
An important parameter while grading an essay is the spelling mistakes. Therefore, number of spelling mistakes in an essay has also been included as a feature for our model. To obtain the results for this, we use the spell checker provider named'enchant'.

4.4 *Domain Information Content*
It is perhaps the most marked feature of our model as it tries to understand the semantics and information content of an essay. To get this feature working, we first figured out the best essay from each set (highest scored essay) then, we pulled out nouns from that essay. These nouns were served as keywords for the particular domain. Then, we fire these words into 'WordNet' (http://wordnet.princeton.edu/) and take out their other equivalent. In this way, for each set, we got a bunch of different words, most relevant to its particular domain. Then, we count the number of domain words in the essayprovided.

5. Approach and Evaluation
Many programs in US cover at one writing. Test like GMAT test; the GRE by ETS, as well as the test conducted by PTE. The written outputs to such item are much more difficult than answers to different questions, and are actually given score by humans. Human raters' scales a writing quality of material helped by a scoring machine that finds the features and writing should have so that it can be rated a certain score. The measure profit of getting scores by humans are as follows (a) Step by step

processing of data provided in the writing (b) compare it the with the writing that are already stored in the dataset it and (c) depending on their examination of the writing; provide the result on the content of the writing. Experienced human teachers are able to mark out and praise a writer's way of imagination and the style of writing he/she possess. A teacher can also examine test giver's aptitude, thinking capability, which includes the argument along with the basic understanding of the facts that are projected in the writing for all its soundness, marks given by humans has its boundaries. To start with, experienced teachers must be assigned the duties. Now, they must be informed the way in which their capability of assigning grades must be certified before including them in the panel of teachers for essay grading.

Finally, they must be observed and can be assigned for more duties to keep intact the truthfulness and acceptability of their rating. In year 2013, more than 695,000 test takers worldwide gave the revised General Test, with each test taker providing to two essays topics, creating a total of 2 million responses. Making use of human in grading of such high quantity, especially in humungous exams like the GRE test, can be pretty exhausting to both mental and physical condition of the teacher, and hence is pretty dangerous. Humans are bound to make certain errors due to insufficient knowledge about every topic that can be tough or even impervious to gauge, which in turn can add unnecessary errors in the output obtained. It is also worth noticing that there has been comparatively a little published research on automated essay grading. Hence, what goes in a rater's mind when he is made to provide scores cannot be exactly determined, particularly under pressure while providing score. This deficiency of knowledge about the human rater can be basis for human scoringcould,inturn,reducethetrustonthe scoresprovidedbyhumans as it is not necessary that they are legitimate. It is because of these problems faced in human grading that different testing platform, such as those for testing or providing result typically makes use of more than two human teachers for every output, and any difference between the raters is made out to get the better result if necessary.

5.1 Automated Scoring

Automated scoring has the capability to give away answers to some of the most basic problems in human essay scoring. In today's systems for scoring based on the computer dependent scoring which is relevant to the output to the quality of text which has been provided in the essays. All these systems works with different variables and these variables can be obtained in various ways most accurately my co joining them using a mathematical approach. Humans on the other hand make relevant decisions based on the various factors and under the effect of different techniques. The effectiveness of automated scoring in respect to human scoring is basically in the amount of precise scoring, maintaining consistency in proving the same set of criteria across all the essays that are obtained while taking a test, this not only is used in grading these all the given essay but also plays an important part in providing a much needed feedback and it can be obtained instantaneously. Computers are not affected by external circumstances nor introspectively added to an essay.

Computers are not impervious to their new group of students. Automated scoring can therefore obtain much greater results and can be much more accurate than the scores provided by humans. Most automated scoring systems can provide a nearly real-time assumption of the performance it has been doing along with the result on different techniques of writing. For example, ETS rater engine can provide a much needed check on various grammar mistakes, different and a lot of new words used, techniques, style of writing, ways of organizing, and developing a new way for written text. On similar notes, Pearson's Intelligent Essay Assessor can give output on 6 different features of written text — the ways in which ideas can be implemented, organizing these ideas logically, different types of symbols and short hands that can be used, providing the fluency in the sentences, different choices of words that can be used, and voice or tone in which the message in conveyed. It is difficult for teachers or human beings in general to provide a feedback this accurate and that too when a lot amount of data has to be graded because a huge number of essays are present in the system. Automation of scoring systems are able to grade essays across various grade scores. Humans, in contrast, are most of the time focuses on a certain grade range along which they can provide grades. Reallocation of a human grader

to a new grade range requires a lot of training along which all the humans which are new to this levels are trained accordingly. Also, there are certain scoring systems that have a capability to provide grades to written text in languages other than English. This capability could aids in the scoring of tests that are supposed to translated in English language in order to rate them and thus the whole effort which goes on in training a large set of data can be used as an alternative method for the better assessment of the test data. Therefore using humans in grading a large volume of data, especially in large-scale assessments like the GRE test, can be pretty exhausting. It is worth making a point that these alternate language capabilities have not been widely studied. Still, they represent a path to improve upon human scoring. Subject-verb agreement will be evaluated by the grammar feature in the e-rater engine, and spelling and capitalization are assessed by the machine.

5.1.1 *Automated Essay Scoring: Applications to Educational Technology*
Intelligent Essay Assessor (IEA) is a software tool which assesses (grades mainly) the essay quality. In IEA Latent Semantic Analysis (LSA) is used, it is mainly used to check the semantic similarity of words, phrases and passages in any given text. Results from constant simulations suggest that LSA does its work with much precision. To grade an essay and assess it, LSA is first trained in a certain specific domain (of topics). Once it is done then the student essays are analyzed by LSA based on their meaning, words used, the content of the essay and the concept explained. Although other essay evaluation techniques mainly focus on grammar, punctuations, and spelling mistakes LSA tries to assess it with more depth by looking into the way of writing, sound of arguments mentioned in the essay, fluency and elegance and mainly the conceptual relation between the whole content written and the topic provided. Evaluating the factors like the style of writing, comprehensibility is a tough task to do since each of them influence the other depending on the usage of words in the essay. But still the scores generated by IEA matched the score generated by the human experts in many topics and the feedbacks provided by both human experts and IEA were not conflicting.

5.1.2 *Automated Essay scoring with e-rater*
In essay writing automated scoring is one essential feature that needs to be introduced soon. Essay writing competitions can be considered as one of the fine examples for constructed response task where students have to write about any given topic. The understanding capability of the AES system is not like us humans. Before generating the essay score a human looks into several important features first such as diction, fluency, whereas AES system tries to find any possible relation between these intrinsic values. In AES system the scoring model is made by analyzing as many essays as possible on a specific topic and pre-scored by human raters. Once this is done then the human scores are predicted using the scores already available and further analyzed. In the end the scores extracted from all the features are combined together which finally gives us a machine-generated score.

5.1.3 *Automated Essay Scoring Using Bayes' Theorem*
By using Bayes' theorem in essay scoring, we expand the classification to a three or four point categorical or nominal scale (for example extensive, essential, partial, unsatisfactory) and introduce a large dataset. The contents in Bayesian essay scoring are mainly features of essay such as (specific words, phrases) and other characteristics of essay like the order in which certain noun-verb pair appears or the order of the concepts explained. Two models from the information science field were taken for creating student essays to test Bayesian essay scoring. The models chosen were improved using 462 essays and two score points. The improved system was tested with 80 new, pre-scored essays with 40 essays in each score group. Used variables were comprised of the two models; use of words, phrases and arguments; two ways for trimming; stemming; and usage of stop words. Although the text classification literature propose urgent improvement on thousands of cases per group, even with the amount of inadequate data used in this study, accuracy over 80% was achieved.

6. Results and Discussion

6.1 Dataset

The dataset used in our work has beenextracted from kaggle.com, it consists of the data from the competition conducted by The Hewlett Foundation. All the essays provided are already human graded and since the number of essays is quite large so they have been divided into 8 sets of essays based on the type of essay.

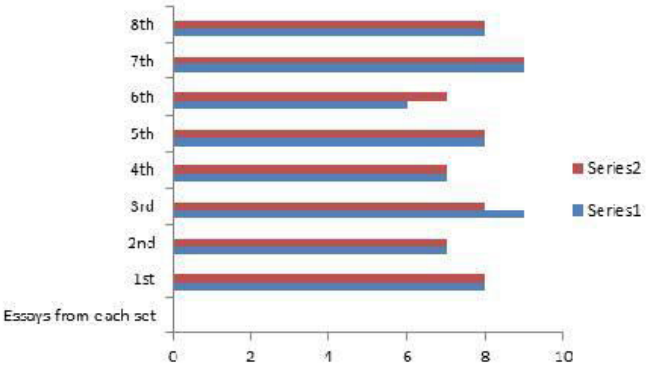
6.2 Technique Used

Our project mainly works on the AutomatedEssay Scoring with e-Rater technique, first of all an essay isprovided as input and then it is compared with the essays of each set once it is done then the essays are compared based on their polarities, words used and the content of essay. The machine finally generates a score for the essay by combining all the results to get a final-score.

6.3 *Comparison between human and machine graded scores*:

In Fig 2 depicts comparison of 8 essays from all the 8 sets graded by both human expert and machine where Series1 represents the human graded scores and Series2 represents the machine graded scores. Reading the graph below it is clearly visible that the difference between both the scores is not much indicating that even the machine is capable of assessing an essay like a human rater.

Figure 2. Comparison between human and machine



7. Conclusion

Automated essay grading is a very vital machinelearning application. It has been studied quite a number of times, using different techniques like latent semantic analysis etc. This current approach tries to model the language features like language fluency, grammatical correctness, domain information content of the essays, and put an effort to fit the best polynomial in the feature space using linear regression with polynomial basis functions. The results which may be obtained will be quite encouraging and legitimate. We might achieve average absolute error that is significantly less than the standard deviation of the human scores. Across all domains used, the proposed approach appears to work very well. The future scope of the given problem can extend in various fields. One such area is to search and model good semantic and syntactic features. For this, various semantic parsers etc. can be used. Other area of focus can be to come up with a betterapproach than linear regression with polynomial basis functions like neural networks.

References

[1] Valenti, S., Neri, F and Cucchiarelli, A. 2003 An Overview of Current Research on Automated Essay Grading
[2] Attali Y and Burstein, J 2006 Automated essay scoring with e-raterï¾½ *The Journal of Technology*https://escholarship.bc.edu/ojs/index.php/jtla/article/view/1650
[3] Kaggle 2012 The Hewlett Foundation: Automated Essay Scoring
[4] Andrew N 2012 CS229 Machine Learning Autumn 2012 Lecture Notes from Stanford University
[5] Lukic A and Acuna V Automated Essay Scoring Rice University
[6] Preston D and Goodman D 2012. Automated Essay Scoring and the Repair of Electronics
[7] Bird, Steven, Edward Loper and Ewan Klein 2009 Natural Language Processing with Python.
[8] Jones, Eric, Travis Oliphant and Pearu Peterson SciPy: Open source scientific tools for Python.
[9] Kelly and Ryan PyEnchant
[10] Pedregosa F, Weiss R and Brucher M2011Scikit-learn : Machine Learning in Python