International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 05 | May 2019 www.irjet.net p-ISSN: 2395-0072

Automated Essay Evaluation using Natural Language Processing

Chahat Sharma[1], Akash Bishnoi[2], Akshay Kr. Sachan[3], Aman Verma[4]

[1]*Assistant Professor, Dept. of Computer Science Engineering, Inderprastha Engineering College, Uttar Pradesh, India*
[2,3,4] *Dept. of Computer Science & Engineering, Inderprastha Engineering College, Uttar Pradesh, India*

Abstract - Automated Essay Scoring (AES) is a great research area to analyze the human expertise. AES is one of the most challenging activities in Natural Language Processing (NLP). It makes use of NLP and Machine Learning (ML) techniques to predict the score and match with the human like grading system. The very first model was PEG (Project Essay Grader) propped by Ellis Page in 1960s. Since then, there have been multiple systems that look at providing either a holistic score to the essay or to score individual attributes of the essay. Examples of a few online systems include Grammarly, Paper-Rater, ETS e-rater and IntelliMetric by Vantage. AES relies sufficient number of essay prompts in order to create a grading model which later is used to evaluate the prompts. With the increasing number of people attempting several exams like GRE, TOEFL, IELTS etc., it'll become quite difficult for the institutes to grade each paper besides the difficulty for humans to focus with a consistent mindset. In this scenario, a person finds it very difficult to grade numerous essays every day within time bounds. This paper aims to overcome and solve the problems faced by human experts by providing them with an interface that can perform their work with the same accuracy. For this, features including Bag of words models, numerical features like sentence count, word count, vocabulary size, perplexity etc. were extracted to grade the essay and achieve maximum accuracy. For this, dataset from Hewlett Foundation was picked to select the best features set, by comparing the accuracy of every possible set. The essay set consist of 4 essay prompts with 12000+ essays. So, to evaluate a large number of essays, such assessments seem expensive and time-consuming task. Even if essays graded by human graders are biased, we as a student feel grateful to a system like that, and that is what motivated us to work on this.

Key Words: Essays, E-Rater, Graduate Management Admission Test, IntelliMetric, Machine Learning, Natural Language Processing, Project Essay Grader.

1. INTRODUCTION

This paper aims to build a machine learning system for automatic scoring of essays written by students. The basic idea is to search for features which can model the attributes like language fluency, vocabulary, structure, organization, content etc. Such a system can have a high utility in many places. A linear regression model is built with polynomial basis function to predict the score of a given essay. The subsequent sections explain the input data, features extraction, detailed approach, results, and future scope of the work. Its objective is to classify a large set of textual entities into a small number of discrete categories, corresponding to the possible grades—for example, the numbers 1 to 10. Therefore, it can be considered a problem of statistical classification. 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, thesaurus etc. 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 parts 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 teacher's point of view. Therefore, creation of such a system that can be accurate in both providing the feedback and grading the performance with consuming a lot of time is required. Computerized scoring has many weak points. In AEEF the scores would be much more detail oriented than the ratings provided by two human graders. 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 grader 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. Literature Survey

2.1 Project Essay Grader (PEG)

Project Essay Grade (PEG) is one of the earliest implemented automated essay grading system. It was developed by Ellis Page in 1966 and others upon the request of the college board. This scoring system was developed to make essay scoring more practical and effective and it relies on style analysis of surface linguistic features of a text block. PEG does not use NLP approach but instead uses proxy measures to predict the intrinsic quality and computer approximations. As no Natural Language Processing (NLP) is used so it does not

take lexical content in account. Proxies refers to the structure of essay which consist of average word length, essay length, number of semicolons or commas, counts of preposition, parts of speech and so on. Proxies are calculated by set of training essays and then transformed to be used in a standard multiple regression along with the essays graded by human to calculate the regression coefficients.

One of the best things about PEG is that it's predicted scores are in close agreement to those of human raters. Second is, this system is computationally tractable which means it can track the errors made by the users. PEG scoring system contains two stages the training stage and scoring stage. In training stage, it is trained on sample of essays and in the scoring stage it makes use of proxies to score the essays. PEG purely relies on a statistical approach based on the assumption that the quality of essays is reflected by the measurable proxies. PEG has been criticized because it does not include semantic features of the essays and only focusing on the structure. Since PEG uses structure of essay to score it was easy to cheat by increasing the length of the essay. It has been modified on several aspects in 1990s.

It achieved a correlation score of 0.87 with human raters. A correlation score/coefficient is a statistical relationship between two variables.

2.2 Intelligent Essay Assessors (IEA)

IEA uses Latent Semantic Analysis (LSA) which is a computational model of human knowledge representation. It is also a method for extracting semantic similarity of words and passages from text. Both aspects are presented elsewhere [6].

It is based on statistical analysis of large amount of text (typically thousands to millions of words). Previous attempts to develop automated essay scoring models have primary focus on style of writing. Focus on content have always remained secondary priority. However, LSA focused on conceptual content, the knowledge conveyed in an essay. *Note: IEA does not make use of NLP techniques.*

In their study [3], they compared the performance against two trained ETS graders. They pick two questions, for question one, the correlation between two graders was 0.86 while correlation of LSA with ETS grader was also 0.86. For questions two, the correlation was 0.87 and 0.86 respectively. Thus, LSA approach was able to perform at the same reliability level as the trained ETS graders.

The biggest advantage of IEA is that, it can flag those essays that are off topic, so that it becomes easy for graders to grade.

2.3 E-Rater

ETS developed e-rater has been used for scoring essays since Feb'1999, developed under the team lead of Jill Burstein. E-rater has been continuously upgrading with newer versions. It generates advisory flag when it encounters anomalous essay & writing samples such as exceeding length, repetition material & off topic response. Scores are reported low & summarized in flag message. E-rater version 12.1 uses 10 features to evaluate the essay such as grammar, mechanics, style, usage, organization, development, word choice, word length, positive features, differential word use. E-rater follows Multiple Linear Regression (MLR) methodology in which the data is split into two sets- training dataset & validation dataset. Training dataset is used to build scoring models and evaluation dataset for evaluating the essay.

Estimated featured weights with human score are maximized on the basis of least-square estimation. The final e-rater scores are scaled to match the distributional mean & standard deviation of e-rater scores & those of human scores. In addition, with Multiple Linear Regression (MLR), Support Vector Machine (SVM), Random Forest (RF), and K- Nearest Neighbors (*k*-NN) can also be used.

Fig 1. Comparision between various Models

The results show that SVM performs better than MLR model in evaluation of human scores. Overall SVM yields highest among all desired methods. It concludes that MLR do not fully give the useful content in the feature variable for prediction. More sophisticated models need to be developed to improve ratings. On the other hand, machine learning models result may not be straight forward as the MLR model. The advantage of the MLR model is that basis for the score is seen in the weight that each feature receives. The SVM algorithm, which employs a set of decision surfaces to separate the essays optimally, works best for our particular datasets.

2.4 IntelliMetric

Graduate Management Admission Council (GMAC®) has long been benefitted from advances in automated essay scoring. It started with ETS® e-rater®, but in January 2006, ACT, Inc. became responsible for GMAT test development and scoring. ACT included IntelliMetric Essay scoring system of vantage learning as part of their initial proposal. It has following approach- Two humans are assigned the work of rating a certain number of prompts. If their reviews differ by more than one score point on a scale of 0 to 6, a third rater is appointed to adjudicates scores. Once a sufficient number of prompts are hand scored, a scoring model is developed and later used for evaluation the prompts.

Vantage used various mathematical models:

Simple word counts- It used a collection of well-developed literature using Bayes Theorem in text classification. The concept is to identify the words or phrases the are more closely associated with essay score.

Probabilistic Modelling- An ACT 2004 evaluation of GMAT AWA prompts founds that 87% candidates obtained scores of 3,4 or 5. This model randomly draws a score from AWA frequency distribution.

3. METHODOLOGY USED

Following features have been included in the model:

3.1 Words Count

For any essay prompt, it is the most important feature for grading an essay. As the number of words increases, the scoring pattern also increases. Above conclusion is obtained after plotting graphs of words count vs score for our dataset.

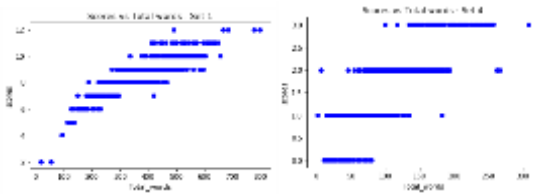

Fig 2. Score vs Total Words for set 1 and set 4

3.2 Sentences Count

An essay is not a single phrase, it consists of many lines supporting writers' points of views. Again, similar trend was recorded, concluding that it also contributes in scoring scheme.

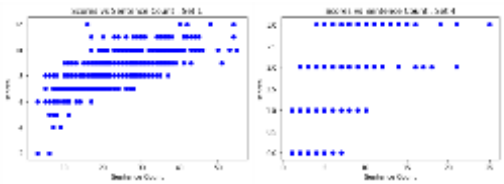

Fig 3. Score vs Sentence count for set 1 and set 4

3.3 Spellings

How good a writer's writing skills also depends on the correct spellings of the words used. This feature also shows a mix trend i.e. positive as well as negative while plotted versus score.

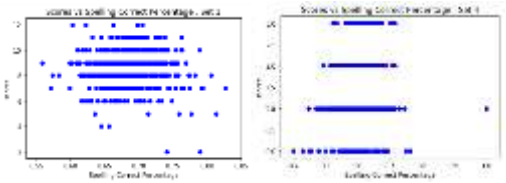

Fig 4. Score vs Correct spelling for set 1 and set 4

2.4 Unique Words

Repetition of words lacks indicates that writer lacks vocabulary. Here, we ignored the stopwords available in NLTK package and words with length less than 3 to get the unique words counts.

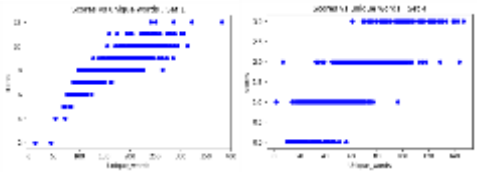

Fig 5. Score vs Unique words for set 1 and set 4

3.5 Grammar Usage

This feature holds the largest weightage after above features when essays are graded manually. In AEG, grammar error doesn't show a variation trend over which a comment can be made. However, we have included it, as it helped in achieving better correlation value irrespective of its much contribution in modelling.

3.6 Words Choice

It is a well-known feature whatever you speak or write, good quality of words presents you better to the second person. We have used NLTK vaden lexicon to predict the sentiment of the words used in the essay (stopwords and words with length less than 3 are ignored).

3.7 Perplexity

Perplexity [8] is a measurement of how well a probability distribution or probability model predicts a sample. It may be used to compare probability models. A low perplexity indicates the probability distribution is good at predicting the samples.

4. RESULT AND DISCUSSION

The dataset [9] has been extracted from Kaggle.com, it consists of data that was used in the competition organized by Hewlett Foundation. The dataset has 4 essay prompts and there are very large number of essays (12000+). Hence, the prompts are divided into 8 essay sets based on their prompts.

Our project mainly works on PEG technique. Important features are extracted from the datasets like total word

count, sentence count, paragraph, correct spellings, parts of speech, perplexity, words choice (positive and neutral words) using word sentiments. Individual Features were plotted against score to identify the trend. All the mentioned features showed a positive trend versus score.

We had trained the selected features using below models to find the correlation between the features and the scores graded by human experts. We got reliable results. The predicted scores are in close agreement with the actual score as shown in table below.

	set 1	set 2	set 3	set 4	set 5	set 6	set 7	set 8
Linear	0.8922	0.7122	0.8127	0.8219	0.8889	0.7253	0.7979	0.6012
best alpha	0.05	0.25	0.05	0.05	0.45	0.05	0.4	1
max lasso	0.8860	0.7322	0.8092	0.8401	0.8938	0.7517	0.8410	0.6358
best alpha	0.15	1	0.4	1	1	1	1	0.35
max ridge	0.8928	0.7262	0.8130	0.8312	0.8921	0.7487	0.8067	0.6195
backward MLR	0.7459	0.6341	0.7766	0.8114	0.8195	0.6760	0.7062	0.4508

Fig 7. Various Models Correlation Matrix

The strength of a correlation	
Value of coefficient R_s (positive or negative)	Meaning
0.00 to 0.19	A very weak correlation
0.20 to 0.39	A weak correlation
0.40 to 0.69	A moderate correlation
0.70 to 0.89	A strong correlation
0.90 to 1.00	A very strong correlation

Fig 8. Meaning of Correlation Score

5. CONCLUSIONS

In sum, we were able to successfully implement a Lasso linear regression model as shown in *Table 1*, using both trivial and nontrivial essay features to vastly improve upon our baseline model. While features like word count appear to have the most correlated relationship with score from a graphical standpoint, we believe that a feature such as perplexity, which actually takes a language model into account, would in the long run be a superior predictor. Namely, we would ideally extend our self-implemented perplexity functionality to the n-gram case, rather than simply using unigrams. With this added capability, we believe our model could achieve even greater Spearman correlation scores. Other features that we believe could improve the effectiveness of the model include parse trees. Parse trees are ordered trees that represent the syntactic structure of a phrase. This linguistic model is, much like perplexity, based on content rather than the "metadata" that are provided by many trivial features. As such, it may prove effective in contributing to the model a more in-depth analysis of the context and construction of sentences, pointing to writing styles that may correlate to higher grades. Finally, we would like to take the prompts of the essays into account. This could be a significant feature for our model, because depending on the type of essay being written— e.g. persuasive, narrative, summary—the organization of the essay could vary, which would then affect how we create our models and which features become more important.

There is certainly room for improvement on our model—namely, the features we just mentioned, as well as many more we have not discussed. However, given the time, resources and scope for this project, we were very pleased with our results. None of us had ever performed NLP before, but we now look forward to apply more statistical methodology to such problems in the future!

6. REFERENCES

[1] Peter W. Foltz, *New Mexico State University*, Darrell Laham, *Knowledge Analysis Technologies*

[2] S. Dikli, "Automated Essay Scoring", Florida State University, Tallahassee (PEG)

[3] Thomas K. Landauer, *University of Colorado,* **"**The Intelligent Essay Assessor: Applications to Educational Technology", *Volume 1, Number 2, October 1999.*

[4] V. Salvatore, F. Neri, and A. Cucchiarelli, "An overview of current research on automated essay grading.", Journal of Information Technology Education: Research 2.1 (2003): 319-330, 2003

[5] L. Hamp Lyons, "The Scope of writing assessment", Elsevier Science Inc, 1075-2935/02 © 2002.

[6] Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990; Landauer & Dumais, 1997; Landauer, Foltz, & Laham, 1998

[7] M. Rudner, V. Garcia, & C. Welch, "An Evaluation of the IntelliMetric™ Essay Scoring System Lawrence", The Journal of Technology, Learning, and Assessment, Volume 4, Number 4 · March 2006

[8] https://en.wikipedia.org/wiki/Perplexity

[9] https://www.kaggle.com/c/asap-aes