INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

**ISSN: 2454-132X**
**Impact factor: 4.295**
***(Volume 5, Issue 2)***
Available online at: www.ijariit.com

# Automated essay scoring using machine learning

*Aarnav Singh*
*aarnav982023@gmail.com*
*SRM Institute of Science and Technology Chennai, Tamil Nadu*

*Daksh Pant*
*dakshkumar_he@srmuniv.edu.in*
*SRM Institute of Science and Technology Chennai, Tamil Nadu*

## ABSTRACT

***Through this project, we aim to build a linear regression model for automated essay grading. Essays form a huge part of certain college applications and are a part of major competitive exams such as GRE, GMAT etcetera. Manually grading the essays can be extremely time consuming and pressurizing for the instructor. To make these tasks simple, we can train an essay grading program to do the grading automatically. The aim of the grader would be to predict the score of the essay with as much accuracy as possible. The machine assigns scores and should be near the human grader's score. A dataset of around 2000 essays was used. The essays were graded on the basis of parameters like word count, coherence, number of long words, correct grammar, etc. Feature selection is used to arrive at the most accurate score prediction.***

***Keywords***— *Automated, Essay, Scoring, Linear, Regression, Model, Quadratic, Weighted, Kappa*

## 1. INTRODUCTION

Automated Essay grading systems are achieving popularity in academic institutions across the world. We can classify these as specialized programs to evaluate written prose. Thus we can save a lot of time for both students and teachers. However, the limitations of computing mean that certain types of prose, such as poetry may never achieve grading par with humans, we estimate that otherwise 90% of all written material in education systems can be evaluated using AES. These Automaton systems score systems by simulation of human behaviour and intelligence in order to relieve the burden of this tedious job called grading from educators leaving them free to educate.

## 2. AES SYSTEMS AND CURRENT STANDARD

Automated essay grading consists of specialized computer programs that assign grades to essays using natural language processing. The basic factors of worth in Automated Essay Grading are cost, accountability, information and technology. However, the use of AES systems in high-stakes examinations has met with severe backlash, with examiners stating that current systems are not accurate enough to warrant a change in the existing system. However, with the use of neural networks, we believe that the use of automation can remove what little human error already occurs and make it efficient enough. The goal of the paper is to prove that AES can be as good, if not better than human graders in select circumstances.

**Table 1: Study on the correlation of electronic grader and three human graders**

| Raters | First human rater | Second human rater | Third human rater | Average human rater | Electronic rater |
|---|---|---|---|---|---|
| First human rater | 1 | .583** | .652** | .857** | .716** |
| Second human rater | .583** | 1 | .641** | .862** | .712** |
| Third human rater | .652** | .641** | 1 | .879** | .890** |
| Average human raters | .857** | .862** | .879** | 1 | .890** |
| Electronic rater | .716** | .712** | .890** | .890** | 1 |

Figure 1, taken from a study labelled "*Automated Essay Scoring Vs Human Scoring: A Reliability Check*" [Ref] helps us to extrapolate the current correlation between human graders and the AES System. From this we can conclude that AES Systems suffer from positive asymmetry, i.e. there are more low marks than high marks. Thus current AES Systems prefer to give average marks rather than award marks inconclusively. On the other hand, sometimes it grades students with a little higher marks than they deserve, i.e. AES to some extent overestimated the quality of the essays. AES Systems have been shown to be better than educators on quantifiable errors such as sentence structure and spelling/grammar errors, the structure of the essay etc. However, other more important properties of the essay such as overall cohesion, the content of the essay are where the systems tend to falter. They can differentiate between both extremes of the spectrum of good and bad but tend to mix up the middling to average content of an essay. Thus to avoid the same we came up with a proposed system of using machine learning techniques over a sufficiently large training set to train our system to be better than the aforementioned system.

The above graph shown in figure 1 is merely the dataset of marks generated by the AES System made into a graph so we can clearly see the bias towards middling marks it suffers from.

 

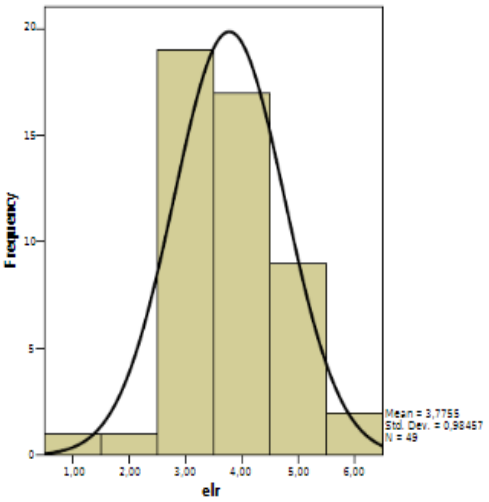


**Fig. 2: Histogram showing AES System grading is asymmetric and heavily leans to the middle ranges**

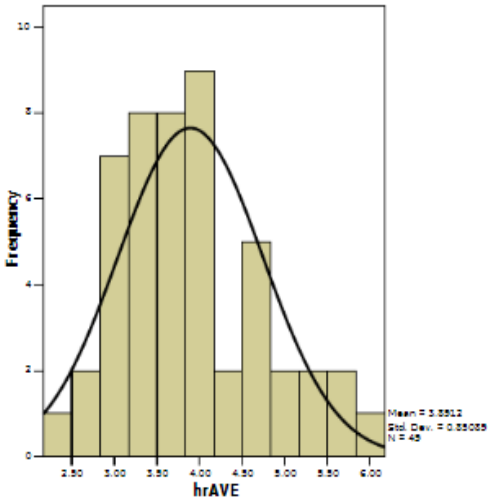


**Fig. 3: Showing average of human graders across the same training set proving positive asymmetric bias of the electronic system**

The Hewlett Foundation sponsored a contest on Kaggle called the Automated Student Assessment Prize. We will be using the same dataset provided during the contest to achieve our goals.

## 3. PROPOSED SYSTEM

As a base for a project, we are working with a set of carefully hand scored essays written in the English language from universities around the world in an attempt to remove any bias from the result that would put in any biases into the AES system that is being built. Our program will evaluate the surface properties of the essay such as linguistics, linear correlation, cohesion, grammar, sentence meaning and overall flow of the essay. In the footsteps of Isaac Parsing and Vincent Ng, we will evaluate the above with the addition of the argument flow of the essay. Then with the above data, we will generate a mathematical data model that relates the quantities with the data generated. For the purposes of our project, we will be using the model of linear regression along with a slight focus on latent semantic analysis (LSA) to find the cohesion and argument flow of an essay.

## 4. MODEL

An illustration of the model used will be provided below. This consists of a total of 4 layers:

1. Input representation layer
2. Memory addressing layer
3. Memory reading layer
4. Output Layer

The Input representation layer is used to generate a vector representation of the response of a student. The Memory addressing layer takes the student's response and divides into random samples and loads the pieces into the memory and assigns a weight to each memory piece. The memory reading layer then extracts the content from the memory depending upon the weights assigned by the previous layer. This produces the output state. The output layer then produces a prediction based upon the output state. We stack the memory reading layer and the addressing layer multiple times to achieve a model extension, which produces a suitable output.

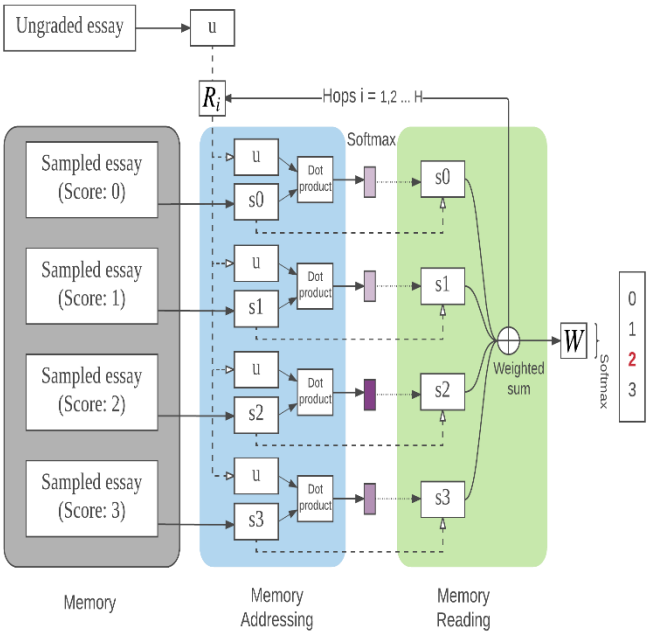


**Fig. 4: Memory networks required for the AES. The range of the score possible is 0-3. A single sample with the same score is selected from the student responses. There are 4 total samples present in the memory**

### 4.1 Input Representation

In our model, we have represented each essay of a student as a vector. Given the number of *n*, we can show each student as $x = \{x_1, x_2, x_3 \ldots, x_n\}$ where *n* is the length of the response. Each word is also treated as a vector $w_i = W_{x_i}$. Thus we can represent the given essay as a vector with position encoding (PE). This is a quick an easy way to store the essay and does not require the need to learn extra parameters.

### 4.2 Memory Addressing

After generating a representation of the essays from the input representation layer, we select a sample from student essay for each score graded with the same score. These samples act as a rubric criterion for all possible scores. Expert knowledge can here be used to choose the most representative sample for each score. All chosen samples are sent into memory as an array of vectors $m_1, m_2, \ldots, m_h$. Here *h* is the total number of samples taken from the represented essays for each score. The weight/importance of each sample $m_i$ is calculated by a dot product among *x* and $m_i$ after a softmax.

### 4.3 Memory Reading

After each vector weight *p* is calculated, the result can be calculated as a weighted sum of each sample in memory in *m*. The formula used is given below:

$$o = \sum_i p_i m_i C^T$$

 

Where $C$ is a $k \times d$ matrix used to keep the reference representation to the feature space. The $C$ matrix is used to train for achieving better performance. The weight vector $p$ shows how much a sample affects the total result.

### 4.4 Multiple Hops

Neural networks secret of success is due to its ability to learn multiple layers of neurons and each layer can transform the representation into a higher level of abstraction. Inspired by this idea, we stacked multiple layers of memory reading and memory addressing one on top of another to handle multiple hop operations.

**Table 2: Selected details of ASAP dataset**

| Set | #Essays | Avg. Len | Max Len | Min score | Max Score |
|---|---|---|---|---|---|
| 1 | 1380 | 350 | 612 | 3 | 23 |
| 2 | 1624 | 350 | 117 | 4 | 5 |
| 3 | 1329 | 150 | 398 | 2 | 9 |
| 4 | 1878 | 150 | 343 | 0 | 7 |
| 5 | 1103 | 150 | 423 | 0 | 6 |
| 6 | 1002 | 150 | 476 | 1 | 6 |
| 7 | 1265 | 250 | 655 | 0 | 28 |

After receiving the output $o$ from the memory reading layer, the ungraded result set is updated with

$$u_2 = Relu(R_1(u + o))$$

Here, the regression model parameters are defined as below:

$R_1$ is a $k \times k$ matrix, $u = xA^T$ and $Relu(y) = \max(0, y)$. Then the memory reading and addressing hop is changed on each consecutive hop using a different $R_k$ on each hop $k$. The updated expression thus becomes:

$$p_i = Softmax(u_j . m_i B)$$

### 4.5 Output Layer

After a fixed number $h$ of hops, we have generated a number of possible solutions. To predict a final output, we use the below formulae to put out a result:

$$\hat{s} = Softmax(u_H W + b)$$

Where $W$ is a $k \times k$ matrix, $r$ is the number of predicted scores and $b$ is the bias value. The amount of result nodes is equal to the length of the score range. Thus by calculating a possible score above distribution of probable scores is how we arrive a final grade value to display for the given sample score. $\hat{s}$ is the predicted score and $s$ is the actual score.

## 5. DATASET

The dataset used in this project was taken from the Kaggle Automated Student Assessment Prize (ASAP) competition sponsored by William and Flora Hewlett Foundation (Hewlett). The dataset was written by students from grade 7 to 10. There are a total of 8 sets. Score range varies across the sets. All essays were graded by at least 2 human graders. The average length of the essays was from 150 to 650 words. The dataset can be found at the following link: https://www.kaggle.com/c/asap-aes/data

## 6. RESULT AND DISCUSSION

In this study, we have developed a generic model for automated essay grading using memory networks. Our model was tested on ASAP dataset and achieved a great score in 6 out of 8 datasets.

This model can thus be used where time is of the essence or to give students a preview of what grade they might achieve before handing in their papers. To improve our model further, we could make a learning vector representation for the assignment. Thus we could eventually associate vector representation to the score.

However, we must concede that our model was tested on only a single dataset, and may have biases that may affect the output. Also, there are several styles of assessments and our model is not optimal for all of them. Future work should thus work on improving the generalizability of the model.

## 7. ACKNOWLEDGEMENTS

We would like to thank S.R.M. University for providing us with a great experience and platform to showcase our skills on this project. Also, we would like to thank our guide Dr Manas Ranjan Prusty for his excellent guidance and help in preparing us for the undertaking of this project paper and instructing in the proper procedure for researching the relevant topics.

## 8. REFERENCES

[1] Hongbo Chen and Ben He. 2013. Automated Essay Scoring by Maximizing Human-Machine Agreement. In EMNLP.

[2] Jeffrey Pennington, Richard Socher and Christopher D Manning. 2014. Glove: Global Vectors for Word Representation.

[3] Sumit Chopra, Jason Weston, Sumit Chopra. 2014. Memory Networks.

[4] Arthur Szlam, Jason Weston and Rob Fergus. 2015. End-to-End Memory Networks. In Advances in Neural Information Processing Systems 28, C Cortes, N D Lawrence, D D Lee, M Sugiyama, and R Garnett (Eds.). Curran Associates.

[5] Kaveh Taghipour and Hwee Tou Ng. 2016. A Neural Approach to Automated Essay Scoring. In ELMNP.

[6] Burstein, Jill (2003). "The E-rater(R) Scoring Engine: Automated Essay Scoring with Natural Language Processing", p. 113. In: Automated Essay Scoring: A Cross-Disciplinary Perspective. Shermis, Mark D., and Jill Burstein, eds. Lawrence Erlbaum Associates.

 