AUTOMATED ESSAY GRADING USING FEATURES SELECTION

Y.Harika[1], I.Sri Latha[2], V.Lohith Sai[3], P.Sai Krishna[4] , M.Suneetha[5]

[1,2,3,4] *IV/IV B.Tech ,*[5]*Professor*
VELAGAPUDI RAMAKRISHNA SIDDHARTHA ENGINEERING COLLEGE
Dept. of Information Technology ,kanuru,A.P,India

---***---

1) **Abstract** - *Automated essay grading has been a research area to maximize Human-Machine agreement for automatic evaluation of textual summaries or essays. With the increasing number of people attempting several exams like GRE, TOEFL, IELTS etc., it'll become quite difficult for each paper to be graded besides the difficulty for humans to focus with a consistent mindset. Now-a-days we require such interfaces to practice in improving efficient writing skills along with their presence as graders in competitive exams. In this scenario a person finds it very difficult to grade numerous essays every day within time bounds. This project aims to solve this problem by making a stable interface that can help the humans for grading the essays.*

This research has been a platform for us to extract features like Bag of words, numerical features like count of sentences and words along with their average lengths, structure and organization; to grade the essay and achieve the maximum accuracy possible. For this we select the best possible features set, by comparing the accuracy of every possible set.

This system basically uses the Sequential forward feature selection algorithm to find out the best possible feature subset. This algorithm starts with empty set and ends with an efficient set. This algorithm has been used because it performs well for small sized dataset. It uses only simple operations, therefore it is easy to implement.

Key Words: Bag of Words, Sequential Forward feature Selection, Features set.

1. INTRODUCTION

Automated Essay Grading has been a research area since early 60's. We'll have to predict the score of an essay that resembles the output given by human readers. It is a difficult task because, we should extract both quantifiable and unquantifiable features like thoughts of writer while inscribing on paper. It may seem to be an easy process for extracting quantifiable features, but to analyze ambiguities in natural language that humans use every day is a challenging task.

Objective of the system is to classify a large set of textual entities into a small number of discrete categories, corresponding to the possible scores—for example, from 1 to 100. Using a training dataset, the artificial environment we create can identify the patterns and tries to predict the next possible output. This project explores text mining approach and it can help with essay scoring.

1.1. Motivation

Many testing programs around the world include at least one essay writing in them. Examples include the GMAT, GRE, as well as the Pearson Test. Some of the strengths of scoring by human graders are that they can (a) consistently score the essay every time, (b) connect it with their prior knowledge, and (c) make a judgment on the quality of the text. Only in 2012, more than 655,000 test takers. [4]

Worldwide took the GRE revised General Test (ETS, 2013), with two essay prompts, producing a total of more than 1.3 million responses.[1] So, involving humans in such assessments evaluation seem to be a laborious task besides it being expensive and time consuming. Besides this if humans are the graders essay score may be biased. So, as a student we think we will be very grateful to get a system like that. And that is what motivated us to work on this.

1.2. Research Goal

Goal is to make the system understand to one of the human languages and complexities in that. We've also tried to find algorithms to check how accurate they work. Further, this has enabled us to know about the automated systems and experiment it with the machine learning algorithm to generate a stable interface to serve our purpose.

2. PROJECT FLOW

2.1. Data

We used a data set provided by Kaggle.com as a part of some online competition. Each essay has one or more human scores. Each essay set has a different grading rubric, from which we've derived a set of rubrics. Each essay is approximately 3-4 paragraphed in length. Some more essays are extracted and graded by human experts.

2.2. Methodology

Few of the Existing Grading systems are Project Essay Grader, Intelligent Essay Assessor; E-Rater etc. They have been focusing to a great extent on different NLP approaches to judge the contents of essay submitted. They've also been considering all the features they desired at the beginning to include in their system. Training essays have been used for matching the content i.e. using them as a dictionary to refer. Hence we've decided to extract more efficient set of features improving score and performance of the system.

2.3. Drawbacks in existing System

Every feature considered either quantifiable or unquantifiable, may not contribute to make the system more efficient. So it can be made better including only a subset of extracted features that are best scoring the given essay. [2]

Besides all these to measure the performance of system they considered 'accuracy' is considered as the only metric.

2.4. System

The proposed system will ensure you the best possible score by using this feature selection algorithm i.e. the "Sequential Forward Feature Selection" algorithm to select the efficient set of features and compute a better result with consistency.

2.5 Algorithm

This process of the algorithm resembles the name itself. It sequentially begins with a null set and proceeds in a forward direction adding one feature after the other. Whenever a new feature is added it also computes the performance and compares it with the previously obtained value. [2]

The better the result obtained the best the newly added feature is considered as and is retained in the set. Otherwise it is removed and the algorithm proceeds further repeating the same process for succeeding features.

The algorithm seems to be as follows: [7]

```
1. Start with the empty set Y0={∅}
2. Select the next best feature as 'x'
       x= max_accuracy[Y,Y + x]
3. Update Y(k+1)=Y(k)+x;
       k=k+1
4. Go to 2
```

When we don't find any more features to consider, the algorithm stops and the obtained result is defined as the efficient feature set among the total set of features considered.

3 PROPOSED SYSTEM

This project begins with 'model creation' using input dataset with essays in it. Values for extracted features are computed and rubrics are designed. Every essay is preprocessed with steps like stop word removal, tokenizing and stemming in it. This also includes 'case folding' i.e. converting into lower case as considered in this particular scenario.

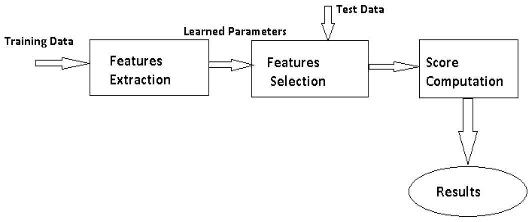


Figure 3.1 System Architecture

Later Input test data is passed and it undergoes the same steps as done to training data. Later computations are done for all features and score is generated as output within a range of 1 to 5. These can be further discretized and graded if needed.

3.2 Data Preprocessing

Data preprocessing is the first step of any data mining approach. Data preprocessing is needed to convert the raw unordered unusable data to structured usable format. Also dataset contains a lot of outliers and other noises that could affect the research in a negative manner. Later on we have made the following modification to our dataset while using them so that we could use them in better way.

3.2.1 Word tokenization

Tokenization is a processing of converting a large text into single pieces of tokens. In our case we have tokenized each essay before using them for feature extraction. We have done word Tokenizing and Sentence Tokenizing. Here is output of a tokenized essay.

3.2.2 Removing Stop Words

Stop words is referred as a collection of words that is mostly present in each and every texts and that are mostly unimportant in research. Removal of such words obtains a better result. These lists contains words like a, an, the etc. And also the punctuations like brackets and other symbols. Though we need a slight modification to enrich the corpus of stop words, it does not include some of the most common words used today. This time the essay contains fewer tokens than the previous one.

3.2.3 Stemming

Stemming is a process that helps to transform the word to its base form or the stem form. For example "am, is, are" all of these can be transformed into base word "be". Car, cars, cars, etc. can be traced down to car. For grammatical reason words are used differently in texts. If we can trace them down to their base form it will help us greatly in processing our data. Among many algorithms to stem the words the porter algorithm is appreciated. We have used this algorithm and reduced the dimensionality of each essays.

4. FEATURES EXTRACTION

After collecting and preprocessing our data we have used different techniques to extract various noticeable features or indicators from the essays so that we can train our model to predict score. We have gathered few features like the following. [6]

Word count after removing stop words: after removing the stop words we have counted the number of words present in each essay.

We have counted the number of sentence each essay contains.

Ratio of words and Sentence: we have calculated the ratio of the words and the sentences and kept them in the feature list

Total number of characters: We have calculated the number of characters present in the essays.

Total number of paragraphs

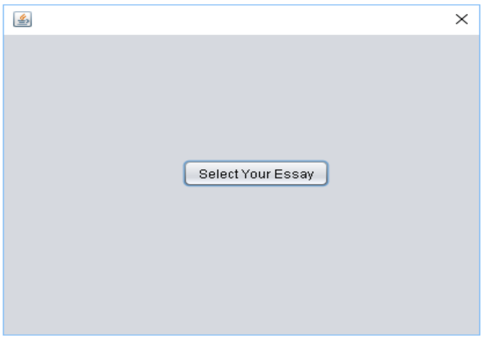
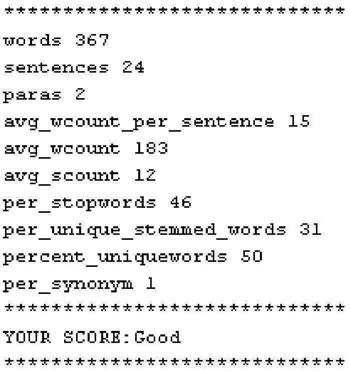


Figure 4.1 Interface to select essay to score

Here we can select a desired text file present anywhere in your system to grade. Then we'll be obtaining the predicted score and few insertions into database takes place i.e. grade for every iteration of subset generated by algorithm as in Figure 4.4. This is the table dumped for visualizing accuracies.



```
*****************************
words 367
sentences 24
paras 2
avg_wcount_per_sentence 15
avg_wcount 183
avg_scount 12
per_stopwords 46
per_unique_stemmed_words 31
percent_uniquewords 50
per_synonym 1
*****************************
YOUR SCORE:Good
*****************************
```

Figure 4.2 Output – Scoring the essay

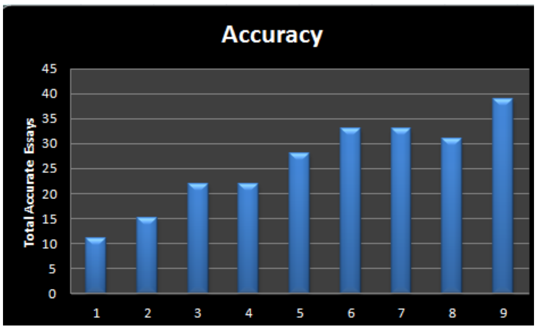
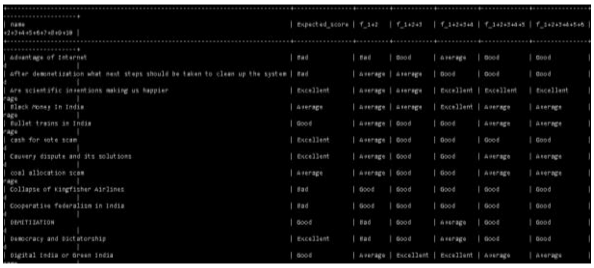


Figure 4.3 Accuracies for every feature



```
+----------------------------------------------------------------------+---------------+---------+-----------+-------------+---------------+-----------------+
---------------------+
| name                                                                 | Expected_score | f_1+2   | f_1+2+3   | f_1+2+3+4   | f_1+2+3+4+5   | f_1+2+3+4+5+6   |
+2+3+4+5+6+7+8+9+10 |
+----------------------------------------------------------------------+---------------+---------+-----------+-------------+---------------+-----------------+
---------------------+
| Advantage of Internet                                                | Bad           | Bad     | Good      | Average     | Good          | Good            |
d                   |
| After demonetization what next steps should be taken to clean up the system | Bad    | Average | Average   | Good        | Good          | Good            |
d                   |
| Are scientific inventions making us happier                          | Excellent     | Average | Average   | Excellent   | Excellent     | Excellent       |
rage                |
| Black Money in India                                                 | Average       | Average | Average   | Excellent   | Average       | Average         |
rage                |
| Bullet trains in India                                               | Good          | Average | Good      | Good        | Average       | Average         |
rage                |
| cash for vote scam                                                   | Excellent     | Average | Good      | Good        | Good          | Good            |
d                   |
| Cauvery dispute and its solutions                                    | Excellent     | Average | Good      | Good        | Average       | Average         |
d                   |
| coal allocation scam                                                 | Average       | Average | Good      | Good        | Average       | Average         |
rage                |
| Collapse of Kingfisher Airlines                                      | Bad           | Good    | Good      | Good        | Good          | Good            |
d                   |
| Cooperative federalism in India                                      | Bad           | Good    | Good      | Good        | Good          | Good            |
d                   |
| DEMITIZATION                                                         | Good          | Bad     | Good      | Average     | Good          | Good            |
d                   |
| Democracy and Dictatorship                                           | Excellent     | Bad     | Good      | Average     | Good          | Good            |
d                   |
| Digital India or Green India                                         | Good          | Average | Excellent | Excellent   | Average       | Average         |
```

Figure 4.4 Database table after algorithm implementation

In Figure 4.4 we find grades inserted for every essay with respect to every feature added at every instance. From this we can find out the count for which the scores already given and predicted are equal to judge the best set of features.

5. RESULTS AND ANALYSIS

From the set of test data we'd, score computations are done for every essay and stored. Later to measure the variations from the Expected Score and Actual Score visually, Database table is dumped into Excel and represented as Figure 5.1.

The above line in the chart indicates the score computed or predicted by the system for every individual testing essay. The bottom line represents the score given by human experts for every essay as in test data. The variation in the points represents a deviation from the score which is to be actually predicted.

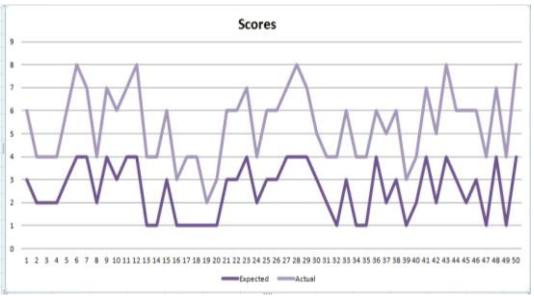


Figure 5.1 Accuracies for every feature

6. CONCLUSION

This research shows that automated essay grading will help writers to know their own level of competency. However, there are many areas of improvement especially with the complexity for languages around the world even in this system some of the critical grammatical errors are quite impossible to register & they have gone unnoticed. Nonetheless, it is quite evident what machine learning can do if the agent can be trained properly with large set of relevant data and reasonably good algorithm will even make the learning easier.

This research suggests that, although statistically significant differences existed between instructor- and AES-grading, it would be quite preferable to use this interface. Consequently, depending on the intent of individual instructors for their chosen assignment, these systems may be more acceptable as a formative, as opposed to summative, assessment of student learning.

ACKNOLEDGMENTS

The present project work is the several days study of the various aspects of the project development. This project work would not have been possible without the support of many people.

We would thank Mrs. Dr. M.Suneetha, for scrupulous guidance and encouragement since the dawn of project work. It lives us immense pleasure in placing on record, greatest sense of gratitude to our project guide. We are highly indebted to her, for consistent and continuous help during the project.

This acknowledgement will be incomplete if we failed to express our deep sense of obligation and reverence to our parents. They motivated us and provided their moral support and encouragement throughout our career.

REFERENCES

[1] Balfour, S. P. (2013), Automated essay scoring and calibrated peer review, Research & Practice in Assessment, 8(1), 40-48.

[2] Farrús, Costa-jussà. (2013), Automatic evaluation using latent semantic analysis, International Review of Research in Open and Distance Learning.

[3] Wang, J., & Brown, M. S. (2008), Automated essay scoring versus human scoring, Contemporary Issues in Technology and Teacher Education.

[4] Shermis, Burstein. (2010), Automated essay scoring: Writing assessment and instruction, International Encyclopedia of Education

[5] Deane, P., Williams, F., Weng, V., & Trapani, C. S. (2013), Automated essay scoring in innovative assessments of writing from sources, The Journal of Writing Assessment

[6] J. Burstein, N. Elliott, & H. Molloy (2016), Automated Writing Evaluation, CALICO Journal Vol. 33.

[7] M. Heilman, F. J. Breyer, F. Williams, D. Klieger, & M. (2015), Automated Analysis of Text in Graduate School Recommendations, ETS Research Report

[8] J. Burstein, J. Tetreault, (2013), The e-rater® Automated Essay Scoring System, Handbook of Automated Essay Scoring: Current Applications and Future Directions.