Automated Essay Evaluation using Chart Parser

Sampada K.S.
Department of Computer Science & Engineering,
RNS Institute of Technology VTU, Bengaluru
Karnataka, India

Anusha S.
Software Engineer
HashedIn by Deloitte
Bengaluru, Karnataka, India

N. Vignesh Karthik
Software Engineer
HashedIn by Deloitte
Bengaluru, Karnataka, India

ABSTRACT
Essays help in assessing academic excellence and linking various ideas with the ability to recall. Evaluating essays manually is a tedious and time consuming job. Automated grading shall reduce the evaluation time and with appropriate training, would generate a realistic and accurate score. We aim to develop an automated essay evaluation system by employing a regressor, fed with features like count of misspelt words, sentences, words, characters, nouns, verbs, adverbs, adjectives, and lemmas. Sentences are checked for grammatical correctness using a custom built parser. The regressors are trained on the enlisted features and then measured for the performance. Various regressors like Linear, Logistic and Random Forest have been employed and observed to select a model with the best performance for use.

General Terms
Natural language Processing, Machine Learning, NLTK POS-tagger, Chart Parser.

Keywords
Natural language Processing, Machine Learning, NLTK POS-tagger, Chart Parser, Linear Regression, Logistic Regression, Random Forest Regression.

.

1. INTRODUCTION
An essay is a short form of literary work, based on a subject matter which often portrays the personal opinion of an author regarding the subject matter. The word essay is derived from the French word essayer, which means "to attempt," or "to try". The Oxford Dictionary describes it as "*a short piece of writing on a particular subject.*"

Essay writing delivers prominence in assessing academic progress, and exhibits abilities to stitch different ideas together [1]. Writing numerous essays and getting them evaluated by an instructor is the best way to improve one's essay writing skills. However, manually grading such essays requires excessive evaluation time. Efficient automated evaluation strategies can significantly reduce the evaluation time and cost. The core interest here is to evaluate and grade essays automatically using Natural Language Processing (NLP) and Machine Learning. Various features are extracted from the corpus to obtain the evaluation metrics using NLP techniques.

A set of common features have been observed and considered for the use across implemented models for essay evaluation. Efforts have been made to identify other significant features as evaluation metrics, to evaluate the essays better.

2. RELATED WORK
Different criterias can be used to evaluate an essay [2]. Three state-of-the-art systems employed are:

- *Intelligent Essay Assessor* considers the number of misspellings, level of diction, context along with sentence redundancy.
- *Criterion* considers typographic errors, verb formation errors, incorrect word use, average word length and missing punctuation.
- *IntelliMetric* considers punctuation, spelling, degree of completeness of sentence, grammar, sentence complexity and vocabulary.

Evaluating an essay considering the nature of language employed on the sentence level instead of an entire essay proved to be more effective [3]. A panel of six professors from a University, evaluated and graded essays manually and inferential statistical tests were performed before and after the training students. Results show a remarkable rise in the scores. The method ensured personalized, accurate feedback to students and helped the university in enhancing quality, but took too long and did not cater to all the needful students.

A tool named "Criterion" was designed to score an essay automatically, and generate personalized feedback. This tool used strategies employing Natural Language Processing and Machine Learning [4]. The E-Reader extracts linguistic features of an essay and uses statistical models to assign holistic scores, trained on a set of 270 essays, manually evaluated on a scale of 1 through 6. A stepwise linear regressor is employed to select features from the essays. The Critique is used to detect errors in grammar and provides feedback. Agreement of the E-Rater score to the human evaluated score is around 97 percent. The application improves the essay writing skills and has low essay processing time, but the system uses simple machine learning models and the aptness of feedback generation is unexplained.

A good essay has the conclusion section clearly demarcated from the rest. To recognize such distinctions, topic dependent and topic independent analysis for a given essay were performed [5]. The training set included a well annotated essay sequence, with the scores evaluated by two evaluators manually. The extracted features were fed onto a C5.0 machine learning model with boosting for the two mentioned implementations. The topic dependent analysis performed slightly better, but the two models have a fair accuracy and good compatibility. The team also plans to make the implementations better and general.

3. METHODOLOGY

3.1 Proposed Methodology
Fig 1. Shows the proposed system. Following sections provide details of each step.

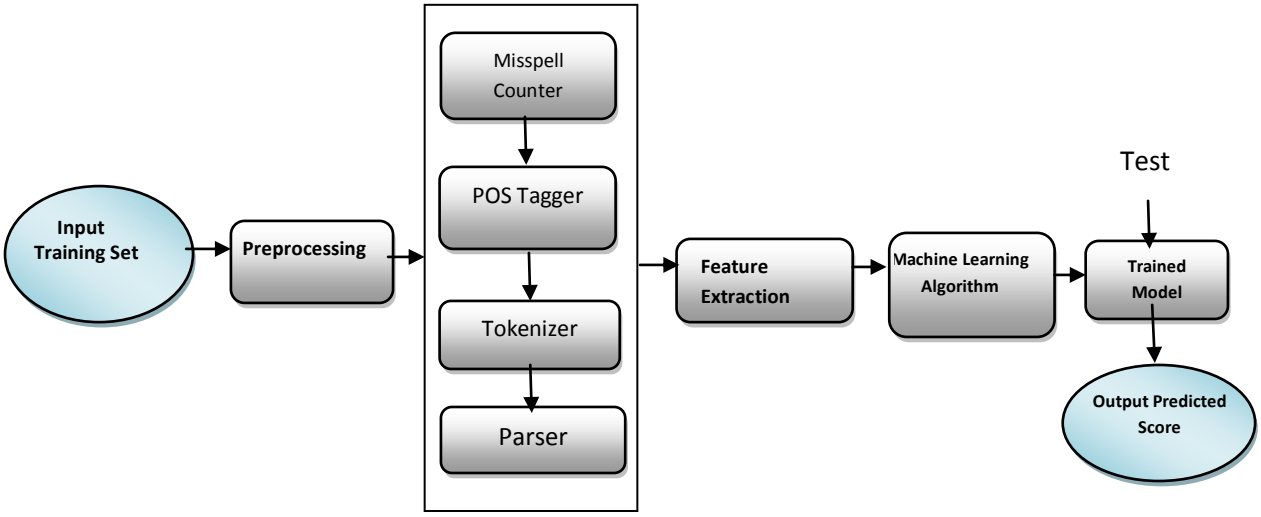

Fig 1: Proposed system

3.2 Procedure

Every essay is split into sentences. The following steps are performed on each sentence.

3.2.1 Preprocessing

Preprocessing done here is to remove the special characters. All the special characters are removed other than full stop('.') or at('@'). Full stop is not removed in order to mark the end of a sentence. In all the essays considered the proper nouns are annotated with '@'(Ex: Name of a person is represented as '@PERSON').

3.2.2 Misspell Counter

A large corpus of correct english words is maintained. Given a word in a sentence, if it's not found in the corpus, it is then considered to be a misspelt word. Since the corpus does not contain any proper nouns, the count of misspelt words is 98% accurate. To further improve the accuracy a SpellChecker package is deployed for the words not present in the corpus.

3.2.3 POS Tagger

Every sentence is fed to a NLTK pos-tagger, also known as the "look-up tagger", assigns tokens to the most frequent morphological labels which appear with it in the training data [6]. This returns a word with its corresponding part of speech. The number of nouns, verbs, adjectives and adverbs along with the number of lemmas are found.

3.2.4 Parser

There are no freely available tools for checking the grammar of a given sentence. Pos_tagger tags the words but does not detect grammatical errors in a sentence, failing to serve the purpose of checking grammatical errors. Hence, the Context Free Grammar rule set was built to detect grammatical errors, working on the lines of a chart parser. The Chart Parser module defines a flexible parser that uses a chart to record hypotheses about syntactic constituents.

The CFG consists of 10 different rulesets. Before feeding a sentence onto the CFG, the length of the sentence is checked. According to global standards, the average number of words per sentence is fifteen. If the count of words per sentence exceeds twenty, it is considered to be grammatically incorrect. Then, it is fed onto any one of these rules, based on the part of speech it consists of. When a sentence is fed to the parser, the parser may:

- return the number of subtrees
- enter an infinite loop.

A sentence is said to be grammatically wrong if the parser enters an infinite loop or the number of subtrees returned is zero. To avoid infinite looping, we have limited the execution time to 15 seconds.

3.3 Machine Learning Models used

The features extracted are further fed onto a regressor. The following regressors are used

3.3.1 Linear Regression

uses data contained in the training set to build a linear model, focusing more on the conditional probability distribution [7]. Linearity of data is directly related to the performance of the model. The following is the equation

$$y = a * x + c$$

Sklearn Linear Regression with default parameters is used.

3.3.2 Random Forest regression

is a decision tree learning algorithm with construction of multiple trees and forests during the training phase, and returns the mean of all individual predictions

$$y = \frac{1}{m}\sum_{j=1}^{m}\sum_{i=1}^{n} w_j(x_i, x')y_i$$

Sklearn RandomForestRegressor with default parameters is used.

3.3.3 Logistic regression

Uses a logistic function to model a dependent variable belonging to one in many classes. Logistic regression is being used in statistical analysis.

$$log(\frac{p}{1-p}) = \alpha + \beta 1(a) + \beta 2(b) + \beta 3(c)$$

3.4 Evaluation criteria

Out of all the features considered, misspelt words and

grammatical errors have the highest weightage. If the count of these values increases, the overall score of the essay decreases. The length of the essay is assumed to be within 300 words, and proper nouns are made available in the required format. The complete essay content has to remain on the specified topic and the diction of the essay is not considered.

The dataset contains the evaluated output as a discrete value whereas the regressor predicts a continuous range of values. To overcome this, the accuracy is calculated by rounding off the obtained value from the regressor. This helps in accurate prediction of the score of the essay which is evaluated.

4. EXPERIMENT ANALYSIS

The data used here is a corpora containing 13k essays, taken from *www.kaggle.com*. The dataset consists of fields: essay id, the complete essay, the scores as evaluated by two different experts, ranging between 1 to 20, 1 being the lowest score and 20 being the highest. The dataset qualifies to be impartial and unbiased. The dataset has been split to 75 percent training and 25 percent test. Essays outside the dataset have not been tested. Experiment has been conducted on a machine with the configuration featuring an 8GB RAM, coupled with Intel Core i5.

The following are the assumptions made during this project:

- Essay length does not exceed 300 words.
- The essay is context based, i.e. the topic of the essay is maintained throughout.
- There are no proper nouns in the essay, and if present, they are converted into a specific format.
- The diction of the essay is not taken into account.

A chart parser is custom built to parse the sentences of the essay grammatically. Average word length of an english sentence is estimated to be 8, and processing these words would require an average of ten seconds on the specified configuration. If the sentence is processed for more than fifteen seconds, or the score is zero, the sentence is considered grammatically wrong.

The following are the attributes considered for evaluation of a given essay which are the scores generated by the grammar parser for correctness, spelling errors, number of sentences, words and characters, average word length, the count of nouns, adjectives, verbs, and adverbs. These attributes are further fed into regressor models such as the Linear Regressor, the Logistic Regressor and the Random Forest Regressor, which have been employed to predict the essay scores. The regressor that returns the most promising result is considered to be the final essay score.

5. RESULTS

5.1 Spell checking error

Returns the count of spelling errors in a sentence.

Table 1: Sample sentences with incorrect spelling count

Sentence	Count of incorrect spelling
How many people have acess to a camputer daily in colege.	3
Homes typically provide areas and facilities for sleeping preparing food eating and hygiene.	0

5.2 Grammar check

Is performed by the custom built chart parser, to identify grammatical errors in a sentence. A successful parse returns the parse time and number of possible subtrees, whereas an unsuccessful parse returns a message

Table 2: Sample sentences with the output obtained from the Grammar Check

Sentences	Obtained output
Homes typically provide areas and facilities for sleeping preparing food eating and hygiene.	Time to parse: 0.348255395889 Number of subtrees: 152064
Slammed the door and left.	Time to parse: 0.013298034667 Number of subtrees: 0
The assignment required students to identify an important character in the novel and explaining how the characters actions influence the plot	Number of words exceeded 20
Read I book.	No output

5.3 Model Performance

All the mentioned regressors have been tested on the same inputs and results have been obtained. The values are tabulated and plotted.

Table 3: Accuracy of the regressors

Regressor	Obtained accuracy (%)
Linear Regression	82.532
Random Forest Regression	88.942
Logistic Regression	85.897

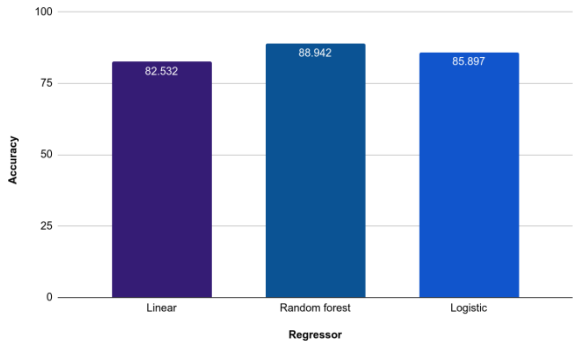

Fig 2: Accuracy plot for regressors

6. CONCLUSION AND FUTURE WORK

6.1 Conclusion

Essays are an important means of improving the thought flow, writing skill sets and on-spot content creation thinking. With the procedure having many advantages, also comes the tedious job of evaluation. Hence, this designed tool can evaluate an essay all by itself and report back to evaluators. Automated Essay Grading is an area where a lot of research has been done in regard to obtaining an accurate model to grade essays automatically. The essays are graded using Natural Language Processing along with Machine Learning techniques. A chart paser has been custom built to evaluate the grammar of the sentence. Due to the consideration of the enlisted features for every essay, the correctness of grammar evaluation and the accuracy of the proposed model is vastly improved, compared to the existing models. By the accuracy obtained, it is safely concluded that the Random Forest Regressor works best for the proposed task.

6.2 Limitations and Future Development

The proposed approach considers various features of an essay to grade it. Though the spelling and grammatical errors have been considered, the result depends on the context and topic of the essay which has not been considered in this release. In future developments, the context of the essays can also be considered as one of the features to train the model. Deep learning techniques can be used to obtain a better overall accuracy. A web based user interface can be developed to enhance the user experience and the ease with which they can utilize the application.

7. REFERENCES

[1] Ramalingam, V. V., Pandian, A., Chetry, P., & Nigam, H. (2018). *Automated Essay Grading using Machine Learning Algorithm. Journal of Physics: Conference Series, 1000, 012030.* doi:10.1088/1742-6596/1000/1/012030

[2] Kakkonen, Tuomo & Sutinen, Erkki, *Evaluation Criteria for Automatic Essay Assessment Systems - There is much more to it than just the correlation*, ICCE 2008 16th International Conference on Computers in Education

[3] Sweedler-Brown, C. O. (1993). *ESL essay evaluation: The influence of sentence-level and rhetorical features. Journal of Second Language Writing, 2(1), 3–17.* doi:10.1016/1060-3743(93)90003-l

[4] Burstein, Jill & Chodorow, M. & Leacock, Claudia. (2003). *CriterionSM: Online essay evaluation: An application for automated evaluation of student essays. IAAI.*

[5] Burstein, J., Marcu, D. A, *Machine Learning Approach for Identification Thesis and Conclusion Statements in Student Essays. Computers and the Humanities* 37, 455–467 (2003) doi:10.1023/A:1025746505971

[6] Giuseppe G. A. Celano*, Gregory Crane, and Saeed Majidi, *Part of Speech Tagging for Ancient Greek*, DOI: 10.1515/opli-2016-0020.

[7] Nelli F. (2018) *Machine Learning with scikit-learn. In: Python Data Analytics*. Apress, Berkeley, CA https://doi.org/10.1007/978-1-4842-3913-1_8.