Automated Essay Grader

Aayush Shah, TwishaVyas, Siddharth Shah, Abhijit Patil

Abstract—Essays are crucial testing tools for assessing academic achievements, integration of ideas and ability to recall, but are expensive and time consuming to grade manually. Manual grading of essays takes up a significant amount of instructor's valuable time, and hence is an expensive process. Automated grading, if proven to match or exceed the reliability of human graders, will significant reduce costs. The purpose of this project is to implement and train machine learning algorithms to automatically asses and grade response. These grades from the automatic grading system should match the human grades consistently. Currently, automated grading is used instead of second graders in some high-stakes applications, and as the only grading scheme in low stakes evaluation.

Index Terms—Automated essay grader; Machine Learning; Natural Language Processing; Linear Regression.

I. INTRODUCTION

NaturalLanguage Processing, NLP, explores how we think and feel, and examines the "inner" language we use to represent our experiences [5]. NLP also studies what we know about human interaction and human achievement, and uses that knowledge to "model" excellence in every walk of life. The theories and techniques of NLP help you to discover what makes some people excel in their lives, and give you the practical tools to do the same. At the same time, Machine Learning gives us the capability to not only discover patterns and trends from increasingly large and diverse datasets but also enables us to automate analysis that has traditionally been done by humans. It also provides confidence levels in the likely success of recommended actions. Essays are the key tools in examining a person's ability of critical thinking, demonstrating understanding of material, structuring and organizing an argument but grading them manually is expensive, tedious and time consuming. Scoring the essays manually does not only consume a lot of time but also the perspective and the scores provided by different graders may be different. The main aim of this project is to make an accurate automated essay grading system which will take into account a list of static and dynamic parameters which would in turn generate a score. Initially, a set of pre-checked essays is fed to the system, for training and testing purpose. A set of features which may serve as proxies for what a human grader might look for are then extracted by the system. A learning model would be developed to learn parameters based on these features. The system then grades the essays using this learning model. The accuracy of the system can be analysed by comparing it with manually graded essays.

Revised Version Manuscript Received on November 18, 2015.

Aayush Shah, Computer Department, Dwarkadas J. Sanghvi College of Engineering, Mumbai, (Maharashtra). India.

TwishaVyas, Computer Department, Dwarkadas J. Sanghvi College of Engineering, Mumbai, (Maharashtra). India.

Siddharth Shah, Computer Department, Dwarkadas J. Sanghvi College of Engineering, Mumbai, (Maharashtra). India.

Prof. AbhijitPatil, Computer Department, Dwarkadas J. Sanghvi College of Engineering, Mumbai, (Maharashtra). India.

II. LITERATURE REVIEW

13,000 essays from Kaggle.com were used as a dataset for training and testing purposes. [4] These essays were divided into 8 sets, each set generated from a separate prompt. Then different approaches are followed on this dataset.

A. Approach 1

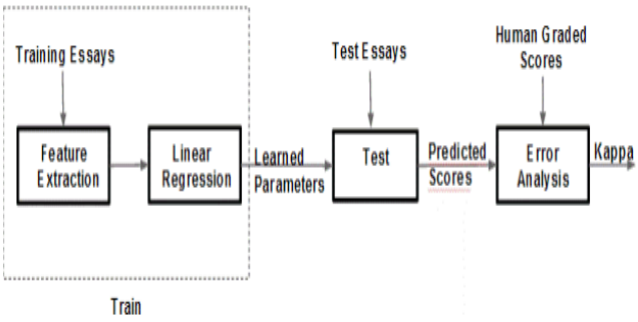
The approach used by ManviMahana et al [2] uses the following features for training the essays:

- Bag of words
- Word count
- Sentence count
- Part of speech count
- Orthography
- Structure and organization

Linear Regression was used to learn parameters based on these features for the training phase. The figure below shows a block level implementation methodology.

A 5-fold cross validation was used in order to guard against over fitting. The quadratic weighted kappa was used as an error metric to measure agreement between predicted and human scores. This model achieved a Kappa score of 0.73 across all essay sets.

Figure 1: Implementation Methodology

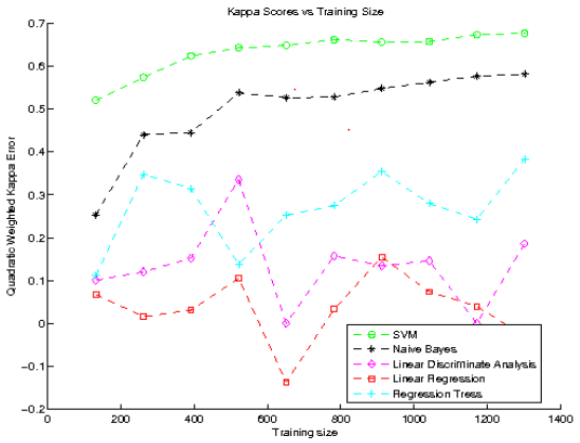


B. Approach 2

The another method used by Shihui Song et al [3] uses features based on a recent ACL paper classifying goodness of scientific New York Times articles. There were five categories of features that were considered and generated for this project for style:

- Inclusion of pronouns
- Beautiful words
- Emotive effectiveness
- Maturity
- Visual Nature

Based on these features, several learning algorithms were evaluated to predict the scores. These algorithms are Linear Regression, Support Vector Machine, Multiclass Linear Discriminative Analysis and Regression Trees. The performance of each algorithm assessed on their Kappa score as



shown below:

Figure 2: Kappa Scores vs Training Size

As shown above, the performance of SVM was the best.

C. Approach 3

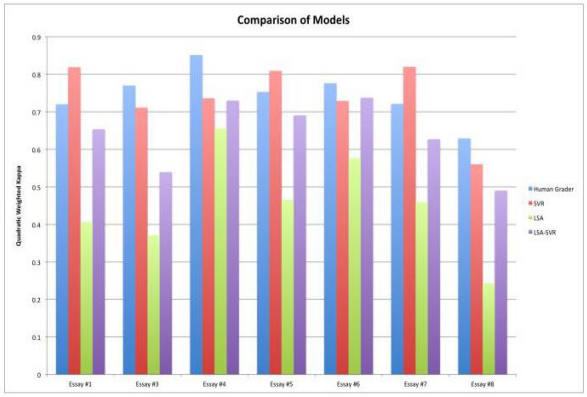
The third approach by Alex Adamson et al [1] approach featurized the essays using common essay grading parameters like character count, word n-grams, word count, sentence count and reduced dimension term-vector. This approach first evaluated the essays using two different algorithms- Support Vector Regression and Latent Semantic Analysis- and then a combination of the two.

While using SVR:

In this technique using the featurized essays, via parameter sweep with C, the choice of kernel and ϵ as free variables, ϵ-SVM is optimized. In some instances, the Quadratic Weighted Kappa of the SVR was higher than the humans - intuitively, the machine agreed more closely with the final human score than the humans agreed with each other.

While using LSA:

This method uses a Singular Value Decomposition to represent the matrix as a product of two orthogonal matrices and a diagonal matrix. By doing this, LSA is able to capture relationships between terms that are not equal, but have similar meanings or concepts. Latent Semantic Analysis was



less successful in producing predictions that agreed with human graders.

Figure 3: Comparison of Models

They also tested a Support Vector Regression that used reduced dimension term-vectors produced by Latent Semantic Analysis.

III. PROPOSED SOLUTION

We plan to take the input data from Kaggle.com. These essays would be divided into 8 sets - each set of essays from a different context - to ensure variability of the domain. Each set of essays is generated from a single prompt. Each essay has been given a score by two human evaluators as well as the average of both the scores. We split this data into two sets: training set and testing set in a ratio of 70:30. Once we have the training and testing data, we can extract features from each of the document and train our model. These are explained in subsequent sections.

At present, our model is using the following set of features to grade the essays:

1. Its Visual Nature: A descriptive sentence awes the reader and prompts their imagination. The source of imagery and meaningfulness used for this data set is derived from the British Natural Corpus where each word has a imagery score between 0 and 999.
2. Inclusion of pronouns: An essay having external references to people, organization or location would be more concrete and engaging. Hence, a count of proper, personal and relative pronouns is maintained.
3. Its beautiful words: Beautiful word choices are thought to increase an essay's elegance, thereby its score. Only words above 5 characters in length are considered beautiful for this project. For a word, the product of its character frequencies is calculated. The lower the product the more complex the word.
4. Its emotive effectiveness: A very dry and emotionless essay is not powerful. The Subjectivity Lexicon from MPQA provides a list of words and their sentiments (positive, negative, neutral, or both) and the strength of those sentiments (MPQA, 2005). The resulting features are proportions of sentiments and strength individually and combined for the entire essay or for given sentiments and strength.
5. Word count and Sentence count: These are very basic features of any text document and do influence the scoring of the document as well. So, to extract these features, we use the "textmining" library (https://pypi.python.org/pypi/textmining/1.0) in python. To get the sentence count, we simply split the document using '.' and thus, count the number of segments obtained.
6. POS Tags: Another crucial set of features for evaluating any piece of writing is the number of words in various syntactic classes like nouns, adverbs, verbs, adjectives etc. These features are crucial for evaluating the quality of content in the essay.
7. Spelling Mistakes: An important parameter while scoring an essay is the spelling mistakes. So, number of spelling mistakes in an essay is also a feature for our model. To get this number, we would use the spell checker provided in python.

8. Domain Information Content: It is perhaps the most crucial feature in our model as it tries to captures the semantics and information content of an essay. To get this feature, we will first take the best essay from each set (highest scored essay) and then pull out the nouns from that essay. These would serve as keywords for the particular domain. Then, we would fire these words into 'Wordnet' and take out their synonyms. In this way, for each set, we would get a bunch of words, most relevant to its domain. Then, we would count the number of domain words in the given essay.

Once we are done with the task of extracting features, the testing phase begins. As the case with training, the testing is also done on essays from all the 8 sets at once. The results obtained in the testing phase are then compared with the human graded scores to obtain the value of Quadratic Weighted Kappa, which would give an assessment of the accuracy of our proposed system.

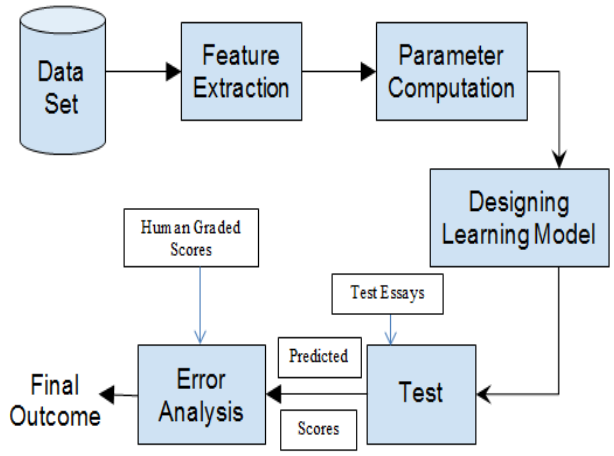


Figure 4: Architectural design

A. Architectural Flow:

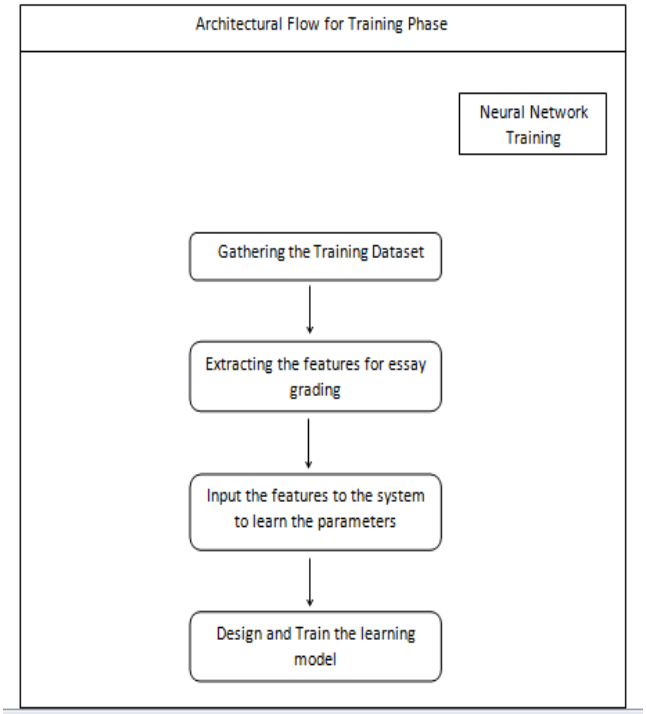


Figure 5: Training Phase Flow

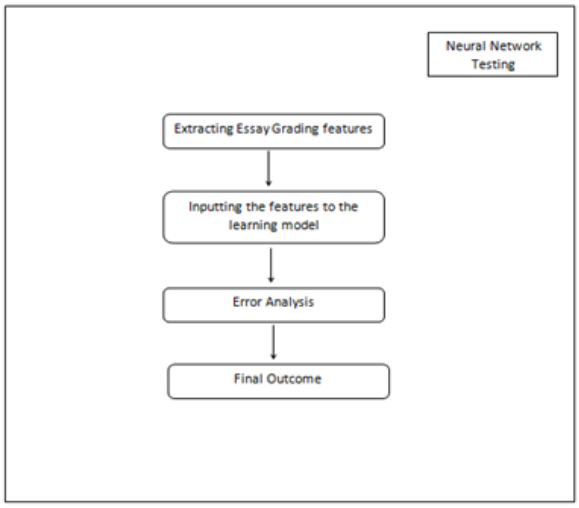


Figure 6: Testing Phase Flow

The above diagrams show the procedures followed in each of the training and testing phase

IV. CONCLUSION & FUTURE SCOPE

Automatic essay grading is a very useful machine learning application. It has been studied quite a number of times, using various techniques like latent semantic analysis, support vector machine and etc. The current approach tries to model the language features like fluency, grammatical correctness, domain information content of the essays and the results obtained seem quite encouraging. Average absolute error, significantly less than the standard deviation of the human scores can be achieved using these approaches.

While these approaches have had fair amount of success in this domain of automated essay grading, it has its drawbacks. Judging by style itself is not apt always. To accurately judge essays we must fathom at an artificially intelligent level both the topic and the essay. As a future prospect, it would be beneficial to explore not only the style by token content, but also on the structural and syntactical style. For this, various semantic parsers etc. can be used. Other area of focus can be to come up with better tools of evaluations like neural networks and etc. Moreover, essays with satirical tone are difficult to understand contextually and using the current features of grading on these kinds of essays could result in inaccurate scores. Hence, this is the area where more research is required to judge the satirical aspect as well.

REFERENCES

1. Adamson, Alex, Andrew Lamb, and Ralph Ma. "Automated Essay Grading." (2014).
2. Mahana, Manvi, Mishel Johns, and Ashwin Apte. "Automated essay grading using machine learning." Mach. Learn. Session, Stanford University (2012).
3. Song, Shihui, and Jason Zhao. "Automated Essay Scoring Using Machine Learning."
4. Preston, Dan, and Danny Goodman. "Automated Essay Scoring and The Repair of Electronics." Technical rep ort, http://snap. stanford. edu/class/cs341-2012/reports/03-Preston_cs341_-_Dan_and_Danny_-_Final. pdf (2012).
5. Natural Language Processing [Online]. Available: https://en.wikipedia.org/wiki/Natural_language_processing