Machine Learning Engineer Nanodegree
Capstone Project
Automated Essay Checking using Machine Learning

Vasu Mistry

May 26, 2017

1 Definition

1.1 Domain Background

Automated essay scoring (AES) has been in the research area of computer science since the early 1966 [1]. Predicting the score of an essay so that the score might seem like it has come from a human reader is a bit daunting task because there are numerous quantified features that have to be extracted from the essay as well as many unquantifiable properties like the perceptions of the writer while writing the essay and his thoughts that he is trying to inscribe on the paper. Therefore, the behavior of the essay inherently noisy, non-stationary and deterministically chaotic. The quantifiable data that can be extracted from an essay is relatively easy for the computer to process rather than processing the ideas or thoughts of the writer in the essay, which may or may not affect the scoring of an essay by a computer.

1.2 Motivation

Personally, I came across this problem during my internship at a start-up. For recruiting content-writers for their website, during the interview phases, a candidate must write a 250 word long essay to demonstrate his skills. At present, this grading is being done manually by the start-up. My goal is to make grading of these essays automatically and I plan to begin my approach through this Capstone Project.

1.3 Problem Statement

The Hewlett Foundation sponsored the Automated Student Assessment Prize on Kaggle - AES Challenge, challenging teams to produce essay evaluation models that best approximate human graders. Many competitive exams try to include maximum number of multiple choice questions since written essays are being graded manually and take up a lot of time to evaluate. While it is a known fact that written essays provide opportunities to challenge the students with more sophisticated measures of ability, the ease of bubbled-answers checking promises a faster evaluation of the student.

The project aims to build a regression model that can take in an essay and automatically output the grade of that essay. Using feature-extraction and Machine Learning algorithms, this task can be automated. The output value will be a continuous value between 2-12.

1.4 Evaluation Metrics

The quadratic weighted kappa score is a measure of agreement of our scores and the human annotator's gold-standard. 0 represents only random agreement between the raters and 1 is full agreement. For N possible essay ratings, an N X N matrix O is constructed where $O_{i,j}$ represents the number of essays receiving grade i from the first grader and j from the second rater. Additionally, a matrix E is constructed the same way, but assuming there is no correlation. The matrices are normalized so that they have the same sum. An N X N matrix w is also calculated where:

$$w_{i,j} = \frac{(i-j)^2}{(N-1)^2} \tag{1}$$

The quadratic weighted kappa is calculated by:

$$\kappa = 1 - \frac{\sum_{i,j} w_{i,j} O_{i,j}}{\sum_{i,j} w_{i,j} E_{i,j}} \tag{2}$$

The Quadratic Weighted Kappa metric typically varies from 0 - only random agreement between raters - to 1 (complete agreement between raters). In the event that there is less agreement between the raters than expected by chance, this metric may go below 0.

2 Analysis

2.1 Datasets

The dataset was provided as a part of the Kaggle - AES Challenge [1]. For this competition, there are eight essay sets with an average length of 150-550 words. The students writing these essays are from Grade 7 to Grade 10. Each essay was graded by 2 or 3 graders and each generated from a single prompt.

For the purpose of this project, Essay sets [1-3] were used to train and test the learning model. The model is capable of acknowledging the difference in scale and outputting the corresponding grade.

Essay Set	Essay Type	Domain	Grade Level	Score Range	Average Length of words	Total
1	Persuassive/Narrative /Expository	-Letter Writing	8	2-12	350	1783
2	Persuassive/Narrative /Expository	- Writing Applications - Language Conventions	10	- 1-6 - 1-4	350	1800
3	Source Dependent Responses	-	10	0-3	150	1726

Table 1: Data-set information and statistics

[1] https://www.kaggle.com/c/asap-aesm

Essays had been anonymized before being released to the public using the Named Entity Recognizer (NER) developed by the Stanford Natural Language Processing group [2]. Replacement IDs of the @ sign followed by words in all capitals were used instead. Name Entities of People, Organizations, Locations, Times/Dates, Numbers, Percents, E-mail Addresses, and Money were replaced.

2.2 Inputs

The data set contains following parameters:

- essay_id: A unique identifier for each individual student essay
- essay_set: The set of a given essay.
- essay: The ascii text of a student's response
- rater1_domain1: Rater 1's domain 1 score
- rater2_domain1: Rater 2's domain 1 score
- domain1_score: Resolved score between the raters.

The following example contains an essay from the data set:

Essay Id	Essay Set	Essay	Rater1_score	Rater2_score	Domain1_score
27	1	Reference: Table 3	4	4	8

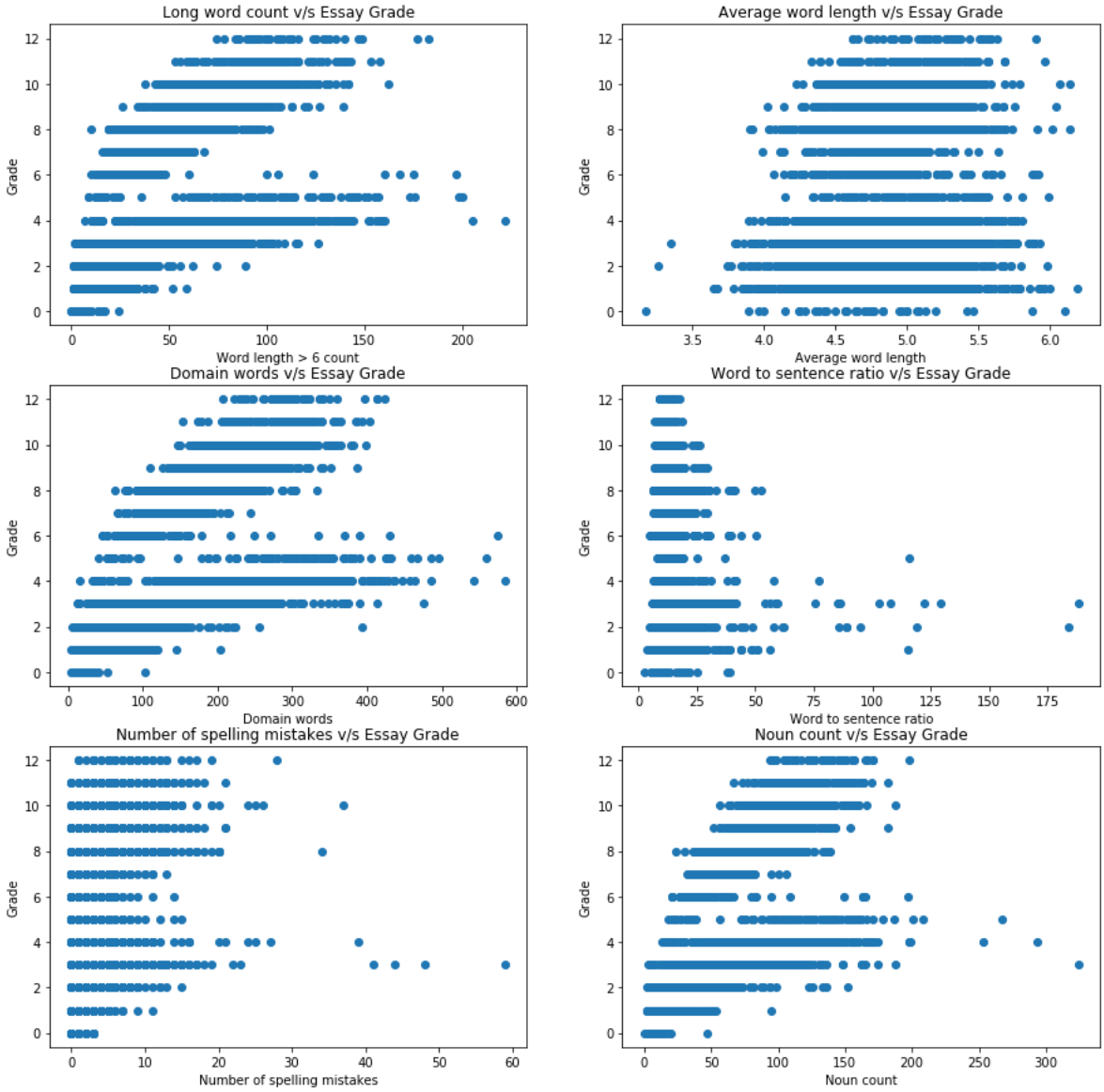
Table 2: Sample data entry

Computers a good because you can get infermation, you can play games, you can get pictures, But when you on the computer you might find something or someone that is bad or is viris. If ther is a vris you might want shut off the computers so it does not get worse. The are websites for kids, like games, there are teen games, there are adult games. Also pictures are bad for kids because most of the time they lead to inapropreit pictures. You should only look up infermation that you need not things like wepons or knifes. Also there are differnt kinds of companies like @CAPS1&t @CAPS2. @CAPS2 is a good place to get computers @CAPS1 so is @CAPS1&t.

Table 3: Sample essay from data set

2.3 Explanatory Visualization

Essay Sets - [1,2,3]



It is inevitable for a good essay to follow a good sentence structure. From the custom features gen-

erated [2], we can clearly see this general trend being followed. Interestingly, it can be observed here that essays with large number of spelling mistakes can receive a higher grade. I would argue that with the advancement of technology in modern spell checker's, the importance for a student to spell a word correctly is given less importance now. The graders clearly consider that the theme/message the student wants to convey in the essay is more important unlike these specific errors.

2.4 Algorithms and Techniques

2.4.1 Linear Regression

An overview of related prior work [3][4][5] indicates that linear regression works well for essay grading applications. In Linear Regression, an output vector denoted y is generated, based on features x extracted from a given essay. Given an input feature vector [3] $x \epsilon R^m$, an output vector [4] $\hat{y} \epsilon R$ using a linear model with a weight of β:

$$\hat{y} = \beta_0 + x^t \beta$$

To learn values for the parameters $\theta = (\beta_0, \beta)$ the sum of squared errors for a training set containing n pairs of essays is minimized and scores,(x_i, y_i) where $x_i \epsilon R^m$ and $y_i \epsilon R$ for $1 \leq i \leq n$:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}=(\beta_0,\boldsymbol{\beta})}{\arg\min} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - (\beta_0 + \boldsymbol{x_i^\intercal \beta})\right)^2$$

2.4.2 Word2vec model

Word2vec is a group of related models that are used to produce word embeddings. These models are shallow, two-layer neural networks that are trained to reconstruct linguistic contexts of words. Word2vec takes as its input a large corpus of text and produces a vector space, typically of several hundred dimensions, with each unique word in the corpus being assigned a corresponding vector in the space [5]. Word vectors are positioned in the vector space such that words that share common contexts in the corpus are located in close proximity to one another in the space [6]

The model will generate a word vector for each unique word in the training set. This vector will be combined with custom heuristic features to create a final vector for each word in the training set.

2.4.2.1 The skip gram model

The training objective of the Skip-gram model is to find word representations that are useful for predicting the surrounding words in a sentence or a document. More formally, given a sequence of training words $w_1, w_2, w_3, ..., w_t$, the objective of the Skip-gram model is to maximize the average log probability:

$$\frac{1}{T} \sum_{t=1}^{T} \sum_{-c \leq j \leq c,\ j \neq 0} log\ p\left(w_{t+j} | w_t\right)$$

[2] Python Notebook : Data_Exploration.ipynb

[3] For building this vector, Google's word2vec model and a few custom generated heuristic features will be used. They are explained in the later sections of this report.

[4] The scores predicted by the learning model

[5] https://en.wikipedia.org/wiki/Word2vec

where c is the size of the training context (which can be a function of the center word w_t). Larger c results in more training examples and thus can lead to a higher accuracy, at the expense of the training time.

2.4.3 Forward feature selection

It is highly unlikely that all of the custom heuristic features generated during the feature extraction will contribute positively towards grading the essay. To eliminate the poor performing features, the greedy forward feature [6] [7] algorithm will be used.

The forward feature selection algorithm works by making changes to the set of features and only keeping the new set if there is an increase in accuracy. Greedy Forward Search works by starting with just one feature and incrementally adding in all the other features. As each feature is added in, the model is evaluated with the feature set and the new feature is only kept if there is a notable increase in accuracy. This is a greedy solution and may not find the absolute optimum feature set, however by looking at which features cause an increase in accuracy it will pick out useful features for the learning model. Also, because the accuracy of the model is evaluated with all the features in a set, this method will pick out features which work well together to achieve a higher accuracy score. Features are not assumed to be independent and so advantages may be gained from looking at their combined effect. [7]

2.5 Benchmark Model

The automated reader developed by the Educational Testing Service, e-Rater, used hundreds of manually defined features. It was trained on 64 different prompts and more than 25,000 essays. Evaluated on the quadratic weighted kappa calculated between the automated scores for the essays and the resolved score for human raters on each set of essays, e-rater could only achieve the kappa score below 0.5 [8]. Hence, I would desire to develop a learning model that has at least a kappa score of 0.5.

3 Methodology

The input does not contain any features, so appropriate features will have to be extracted. An overview of related prior work [3][4][5] indicates that linear regression works well for essay grading applications, hence I will use Linear Regression for the learning model. Since, no validation test essay set was provided in the contest, a 5-fold cross validation will be done to train and test the learning model. This will help to guard against overfitting[3]

4 Project Design

4.1 Programming language and Libraries

- Python 2.7
- Scikit-Learn Open source machine learning library for Python

[6] https://en.wikipedia.org/wiki/Feature$_s$*election*
[7] https://www.cs.cmu.edu/ kdeng/thesis/feature.pdf

- NLTK (Natural Language Processing Toolkit) Open source Natural Language Processing Library [8]
- Grammar-Check Open source library for checking grammatical and spelling mistakes
- NumPy and Pandas - A fundamental package for scientific computing in Python. They provide advanced data structures which are highly useful for data analysis.
- Textmining A Python library for performing statistical text mining.

4.2 Methodology

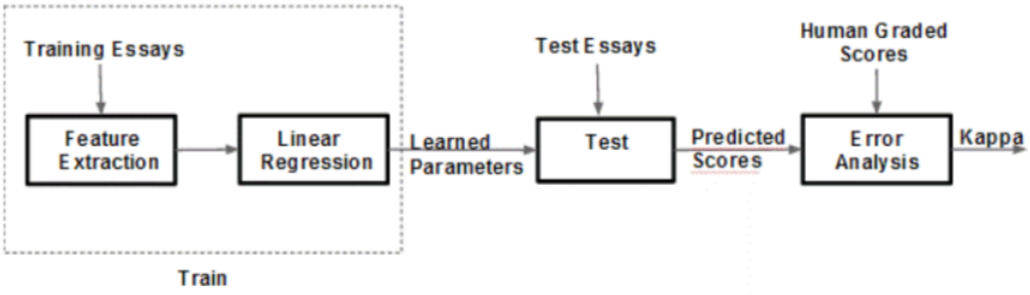


Figure 1: Implementation Methodology.

We hypothesize that a good prediction for an essay score would involve a range of feature types such as language fluency and dexterity, diction and vocabulary, structure and organization, orthography and content. A good model would incorporate features from each of these areas to arrive at a good prediction [3].

4.3 Step Wise approach

1. **Preprocessing:**
   The data was already preprocessed by the organisation using the Stanford NLP tools [9]. The named entities were marked with *""@Text"*. Necessary steps will be taken to ignore these tagged words during the feature extraction (especially spelling mistake checking) stage.

2. **Feature extraction:**

   - Heuristic features
     Several heuristic features that are likely to contribute to a good essay will be generated. Some of the heuristic features are: word count, long word count, average word length per essay, quotation mark count etc.
   - Spelling and Grammatical features
     It is likely for a student to make grammatical and spelling errors, using the open source libraries mentioned in section 7.1 these features will be generated.
   - Part of Speech (POS) tags
     A **count** for most regular POS tags will be used, for example, count of Proper Noun count, Adjective count, Adverb count etc.

[8] Prerequisite: Java
[9] https://nlp.stanford.edu/software/

- Other features
  Several other features like domain words (number of words that relevant to the domain of the essay), punctuation count, word to sentence ratio etc will also be generated.

3. **Training and Cross Validation:**
   The features extracted will be fed to a Linear Regression training model and a 5 fold cross validation strategy will be used to guard against overfitting.

4. **Forward feature selection:**
   It is highly unlikely that all of the features generated during the feature extraction will contribute towards grading the essay. To eliminate the poor performing features, a greedy forward feature [10] [11] algorithm will be used.

5. **Final training and testing:**
   After obtaining a list of selected features from the above step, the learning model will be trained and tested. The predicted values will be checked with the evaluation metrics to verify if the learning model has performed better than the benchmark set or not.

References

[1] Md. Haider Ali Annajiat Alim Rasel Arshad Arafat, Mohammed Raihanuzzaman. *Automated Essay Grading with Recommendation*. 04 2016.

[2] Trond Grenager Jenny Rose Finkel and Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling. *Proc. of ACL*, pages 363–370, 2005.

[3] Ashwin Apte Manvi Mahana, Mishel Johns. *Automated Essay Grading Using Machine Learning*. 12 2012.

[4] Neri F. Cucchiarelli A. Valenti, S. *An Overview of Current Research on Automated Essay Grading,*. 2003.

[5] Burstein J. Attali, Y. Automated essay scoring with e-rater® v.2. *The Journal of Technology, Learning, and Assessment*, 2006.

[6] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc., 2013. URL http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf.

[7] Matthew Shardlow. *An analysis of feature selection techniques*. URL https://studentnet.cs.manchester.ac.uk/pgt/COMP61011/goodProjects/Shardlow.pdf.

[8] et al. Foltz, PeterW. *Implementation and applications of the Intelligent Essay Assessor*. Routledge Handbooks, 2013.

[10] https://en.wikipedia.org/wiki/Feature_*selection*
[11] https://www.cs.cmu.edu/ kdeng/thesis/feature.pdf