An empirical analysis of machine learning models for automated essay grading

Deva Surya Vivek Madala [1] , **Ayushree Gangal** [1] , **Shreyash Krishna** [1] , **Anjali Goyal** [1] , **Ashish Sureka** [Corresp. 1]

[1] Computer Science, Ashoka University, Sonepat, Haryana, India

Corresponding Author: Ashish Sureka
Email address: ashish.sureka@ashoka.edu.in

Background. Automated Essay Scoring (AES) is an area which falls at the intersection of computing and linguistics. AES systems conduct a linguistic analysis of a given essay or prose and then estimates the writing skill or the essay quality in the form a numeric score or a letter grade. AES systems are useful for the school, university and testing company community for efficiently and effectively scaling the task of grading a large number of essays.

Methods. We propose an approach for automatically grading a given essay based on 9 surface level and deep linguistic features, 2 feature selection and ranking techniques and 4 text classification algorithms. We conduct a series of experiments on publicly available manually graded and annotated essay data and demonstrate the effectiveness of our approach. We investigate the performance of two different features selection techniques (1) RELIEF (2) Correlation-based Feature Subset Selection (CFS) with three different machine learning classifiers (kNN, SVM and Linear Regression). We also apply feature normalization and scaling.

Results. Our results indicate that features like world count with respect to the world limit, appropriate use of vocabulary, relevance of the terms in the essay with the given topic and coherency between sentences and paragraphs are good predictors of essay score. Our analysis reveals that not all features are equally important and few features are more relevant and better correlated with respect to the target class. We conduct experiments with k-nearest neighbour, logistic regression and support vector machine based classifiers. Our results on 4075 essays across multiple topics and grade score range are encouraging with an accuracy of 73% to 93%.

Discussion. Our experiments and approach are based on Grade 7 to Grade 10 essays which can be generalized to essays from other grades and level after doing context specific customization. Few features are more relevant and important than other features and it is interplay or combination of multiple feature values which determines the final score. We observe that different classifiers result in difference accuracy.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3518v1 | CC BY 4.0 Open Access | rec: 9 Jan 2018, publ: 9 Jan 2018

An Empirical Analysis of Machine Learning Models for Automated Essay Grading

Deva Surya Vivek Madala[1], Ayushree Gangal[1], Shreyash Krishna[1], Anjali Goyal[1], and Ashish Sureka[1]

[1]**Ashoka University, Haryana, India**

Corresponding author:
Ashish Sureka[1]

Email address: ashish.sureka@ashoka.edu.in

ABSTRACT

Background. Automated Essay Scoring (AES) is an area which falls at the intersection of computing and linguistics. AES systems conduct a linguistic analysis of a given essay or prose and then estimates the writing skill or the essay quality in the form a numeric score or a letter grade. AES systems are useful for the school, university and testing company community for efficiently and effectively scaling the task of grading a large number of essays.
Methods. We propose an approach for automatically grading a given essay based on 9 surface level and deep linguistic features, 2 feature selection and ranking techniques and 4 text classification algorithms. We conduct a series of experiments on publicly available manually graded and annotated essay data and demonstrate the effectiveness of our approach. We investigate the performance of two different features selection techniques (1) RELIEF (2) Correlation-based Feature Subset Selection (CFS) with three different machine learning classifiers (kNN, SVM and Linear Regression). We also apply feature normalization and scaling.
Results. Our results indicate that features like world count with respect to the world limit, appropriate use of vocabulary, relevance of the terms in the essay with the given topic and coherency between sentences and paragraphs are good predictors of essay score. Our analysis reveals that not all features are equally important and few features are more relevant and better correlated with respect to the target class. We conduct experiments with k-nearest neighbour, logistic regression and support vector machine based classifiers. Our results on 4075 essays across multiple topics and grade score range are encouraging with an accuracy of 73% to 93%.
Discussion. Our experiments and approach are based on Grade 7 to Grade 10 essays which can be generalized to essays from other grades and level after doing context specific customization. Few features are more relevant and important than other features and it is interplay or combination of multiple feature values which determines the final score. We observe that different classifiers result in difference accuracy.

1 INTRODUCTION

1.1 Research Motivation and Aim

Automated Essay Grading or Scoring (AEG or AES) consists of automatically evaluating the score or grade of a written essay (Cummins et al., 2016)(Dong and Zhang, 2016) (Balfour, 2013) (Chen et al., 2010). AES systems are motivated by the need to develop solutions for assisting teachers in grading essays in an efficient and effective manner. AES systems are also useful for students to understand issues in their writing by receiving a quick feedback from a system rather than waiting for inputs from a teacher. Accurate and reliable AES systems are needed by schools, universities and testing companies to be able to manage the grading of essays by large number of students. One of the main technical challenges in building an AES system is to be able to achieve an output which is in agreement with a human evaluator. AES systems has attracted the attention of several researchers and several solution approaches have been proposed (Cummins et al., 2016)(Dong and Zhang, 2016) (Balfour, 2013) (Chen et al., 2010). However, AES is still not a fully solved problem and we believe more research and alternative novel approaches are needed to further enhance the state-of-the-art. Our research work presented in this paper is motivated by the need to conduct experiment on the effectiveness of several linguistic features and variables for

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3518v1 | CC BY 4.0 Open Access | rec: 9 Jan 2018, publ: 9 Jan 2018

estimating the score of an essay written primarily by middle school students. While the framework and methodology presented in our work can be generalized, our focus is on grading essays of school students from Grade 7 to Grade 10. Our motivation is to investigate whether writing skills can be assessed by automatically checking aspects such as richness in vocabulary, word count with respect to the prescribed limit, semantic similarity of the terms in essay with the topic of the prose, usage of active and passive voice, semantic similarity and coherence of terms in the essay body, spelling errors, usage of tense, grammatical errors and sentence lengths.

There are several research gaps and open research questions in the area of automated essay scoring and grading. One the research questions pertains to identification of relevant and important textual features which can be used to predict the writing skill of the student and quality of the essay. Our aim is to investigate 9 different features for automated essay scoring task. Few of the features are surface level and few require a deeper natural language processing. Our aim is to investigate the effectiveness of 9 features in which few are positively correlated to quality and few are negatively correlated. Conducting experiments on 9 surface level and deep features, positively and negatively correlated features with the score is one of the unique contributions of our work. Our aim is to understand whether our proposed 9 features can be considered as proxies to determine the quality of a student essay at the middle school level. Information retrieval, natural language processing and machine learning have applied several techniques and computational tools (refer to the Literature Survey and Related Work Section of this paper) for computing the score of a given essay. Machine learning is a vast area consisting of several algorithms and methods. Our research aim is to examine the performance of algorithms (and combination of algorithms in a data processing pipeline) which are relatively unexplored. Our aim is to investigate the performance of two different features selection techniques (1) RELIEF (2) Correlation-based Feature Subset Selection (CFS) with three different machine learning classifiers (kNN, SVM and Linear Regression). The main research contributions of our work in context to the existing work on AES is the application of 9 surface level and deep linguistic features, 2 feature selection techniques, 3 classification algorithms on 3 real-world manually annotated publicly available dataset for the task of automatically grading essays. We conduct a series of experiments and conduct a focused and in-depth analysis of our proposed solution approach.

1.2 Literature Survey and Related work

Automated essay scoring and assessment is an important and a technically challenging task and hence attracted the attention of several researchers in the area of machine learning and information retrieval. In this Section, we present several closely related work to our research presented in this paper. Cummins et al. present a constrained multi-task learning approach for automated essay scoring (Cummins et al., 2016). They develop a ranking model using several features such as essay length, grammatical relations, max-word length and min-sentence length and part-of-speech counts (Cummins et al., 2016). Dong et al. propose an approach based on neural networks for automatically learning features for the task of automated essay scoring (Dong and Zhang, 2016). They compare the effectiveness of automatically induced features with handcrafted features and conclude that automatically induced features results in good performance (Dong and Zhang, 2016). Yannakoudakkis et al. use rank preference learning to explicitly model the grade relation between answer scripts (Yannakoudakis et al., 2011). The rank reference system achieves performance close to the upper bound of the task of grading ESOL texts (Yannakoudakis et al., 2011).

Ross et al. use machine learning SIDE program to automatically evaluate the accuracy of 565 students' written explanation of evolutionary change (Nehm et al., 2012). Using Kappa inter-rater agreement between the program and human rater, the SIDE performance was found most effective when scoring models were built using individual item level (Nehm et al., 2012). In subject specific essays such as Life Sciences, Ross et al. investigate the impact of misspelled words on scoring accuracy of a model (Ha and Nehm, 2016). They establish that misspelled words have a greater impact on naive ideas as compared to key concepts and false positive feedback (Ha and Nehm, 2016). Balfour et al. compares human based UCLA's calibrated peer review (CPR) with the Automated Scoring System(AES) (Balfour, 2013). They reason that for several types of essays, AES gives immediate feedback while CPR is better applied to train students with Evaluation Skills (Balfour, 2013). Rudner et. al provide two-part evaluation of the Intellimetric scoring system for Analytical Writing Assessment in GMAT (Rudner et al., 2006). Using a weighted probability model, they infer Pearson r-correlations of agreement between human raters and the

IntelliMetric system averaged .83 in both evaluations (Rudner et al., 2006).

Bin et al made use of the kNN algorithm to categorise essays (Bin et al., 2008). They first converted essays into vectors in a VSM and after filtering out the stop words, employed the Information Gain technique to conduct feature selection. They observed that the best results are given by k = 3 and k = 5 and that words and phrases give poorer results as compared to arguments (Bin et al., 2008). Kakkonen et al employed Probabilistic Latent Semantic Analysis (PLSA) and Latent Semantic Analysis (LSA) techniques to grade essays written in Finnish by setting the similarity metric as the cosine of the angle (Kakkonen et al., 2005). They concluded that although LSA's and PLSA's performances were similar, the former performed marginally better than the latter (Kakkonen et al., 2005). McNamara et al. made use of three techniques for grading essays: Coh-Matrix, the Writing Analysis Tool and the Linguistic Inquiry and Word Count (McNamara et al., 2015). They carried out correlations between the variables reported by the three techniques and then employed various filtering methods to reduce the number of variables from 320 to 140. Next, a discriminant function analysis (DFA) model was used whose accuracy was judged based on: chi-square, Pearson *r*, Cohen's Kappa, exact accuracy, and adjacent accuracy (McNamara et al., 2015). Chen et al. propose an unsupervised approach to essay grading by focusing on the similarity between essays rather than assuming any prior score information (Chen et al., 2010). They employed the voting algorithm and concluded by observing the limited scope of the bag of words model, especially in the domain of creative writing (Chen et al., 2010).

2 MATERIALS AND METHOD

2.1 Research Framework and Solution Approach

Figure 1 shows the research framework and solution approach for our automatic essay grading system. As shown in the Figure 1, the framework consists of 3 sets of data (separate training and test dataset for each of the 3 sets), data scaling and normalization technique, two feature selection techniques (RELIEF and CFS), three classification techniques (kNN, linear regression, SVM) and performance evaluation metrics. We use the publicly available dataset from Kaggle[1] so that our experiments can be easily replicated and can be used for comparison with other approaches. The dataset consists of several essays having an average length 150 to 550 words. The essays are written by middle school students from Grade 7 to Grade 10. We propose and implement 9 features. The 9 features are: Vocabulary, Word Count Limit Ratio, Semantic Similarity Topic Essay, Voice, Semantic Similarity Essay, Spell Errors, Tense, Grammatical Errors and Long Sentences. We apply feature scaling and normalization before providing it as input to the machine learning algorithm. We need to rescale the values as the scale and range for all the features are different. Following is the brief description of the proposed 9 features.

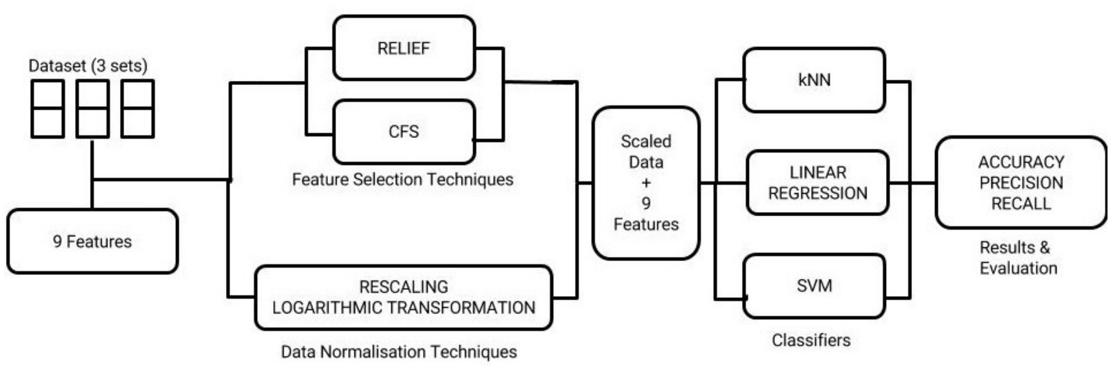

Figure 1. High Level Solution Approach and Research Framework Diagram

Word Count Ratio This feature calculates the ratio of the word count of the given essay with respect to the specified word limit. Our objective is to measure how far the given essay is from the specified word limit in terms of the extent to which the given essay being either too many or too few words.

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

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3518v1 | CC BY 4.0 Open Access | rec: 9 Jan 2018, publ: 9 Jan 2018

This feature assumes equal weightage for equal number of words above or below the word limit. This feature uses Python library textstat[2] to tokenise and count the number of words in the document. The score for this feature is calculated as : (1-WC/WL) where WC represents the word count of the given essay and WL represents the world limit provided in the essay guideline. Subtracting the ratio from 1 is a way of normalising the score (equivalent to taking absolute value of the ratio).

Sentence Length Research shows that very long sentences are hard to comprehend and hence less effective and less coherent due to their high verbosity. Presence of many long sentences negatively impact the final grade of the essay. This feature computes the number of long sentences. We use our word count feature discussed above. We use the Python NLTK library[3] to tokenise the text into sentences and count the total number of sentences. The score for this feature is calculated by dividing the number of sentences having 15 or more words by the total number of sentences in the essay. A large ratio implies that an average sentence of the essay is long.

Voice of the Essay Essay graders recommend that any piece of writing or prose be in active voice rather than passive voice for a better coherency and comprehension. This feature evaluates to what extent sentences in the given essay has been consistently written in active or passive voice. For computing the value of this feature, we use SpaCy Python toolkit[4] to identify the voice of a sentence by analysing the structure of the sentence. For example, in active voice, the subject performs the active verb's action whereas in passive voice, the subject gets acted upon by the verb's action (which is no longer active). The score for this feature is calculated by dividing the number of sentences written in active voice by the total number of sentences in the given essay. A large ratio suggests that an average sentence in the essay is written in active voice.

Tense of the Essay Essay graders and educators recommend that a good piece of writing should be written consistently in the same tense (regardless of the choice of tense). Mixing different tenses may result in a negative impact on the final grade as it makes the essay difficult to comprehend and understand. This feature uses the NLTK Python library to identify different parts of speech (such as verbs, nouns, adjectives) and focusses mainly on verbs. The score for this feature is calculated by first determining what is the dominant tense verb in the essay? Further calculation is done by dividing the number of such verbs by the total number of verbs. A large ratio implies that there is one dominant tense in the essay which is positively correlated with good writing skills and score.

Spell Check It is natural that a good piece of writing minimises the number of spelling errors. This feature first tokenises the text into words and then uses Enchant spell checking library[5] to look up the spelling of these words and returns a count of the number of spelling errors occurring in the document. The score for this feature is calculated by dividing the number of spelling errors by the total number of words in the document. A large ratio suggests a high number of spelling mistakes, which has a negative influence and correlation with the essay score.

Grammatical Errors Similar to spelling errors, it is natural that grammatical errors reduces the essay quality and comprehension. We compute the proportion of grammatical errors in the document by using language-check module in Python[6]. For each sentence, language check checks whether the sentence follows certain grammatical rules or not. The score for this feature is calculated by dividing the number of grammatical errors by the total number of words in the document. A large ratio implies a large number of grammatical errors. A large ration is negatively correlated with the essay score.

Vocabulary Using a rich vocabulary and appropriate vocabulary usage is a good indicator of writing quality. Good organization of ideas and creating a syntactic variety requires good vocabulary. This feature employs a bag of words model. We use the NLTK Python library to first tokenise the text into words and then remove all the stop words and returns the count of unique words. The score for this feature is calculated by dividing the number of unique words by the word limit. The rationale

[2] `https://pypi.python.org/pypi/textstat/0.1.4`
[3] `http://www.nltk.org/`
[4] `https://spacy.io/`
[5] `https://www.abisource.com/projects/enchant/`
[6] `https://bitbucket.org/spirit/language_tool`

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3518v1 | CC BY 4.0 Open Access | rec: 9 Jan 2018, publ: 9 Jan 2018

behind using word limit (rather than word count) is to take care of cases where the ratio may be high owing to the fact that the essay had very few words (essays which are much shorter than the prescribed word limit). A large feature value or ratio implies good use of vocabulary only in cases where the essay is of a sufficient length this influencing the final grade in a positive manner.

Semantic Similarity (two features) We propose two features on semantic similarity and coherency. Semantic similarity of the essay content with the topic and semantic similarity and coherency between terms in the essay body. These two separate features determine to what extent the essay is coherent as well as relevant to the given topic. The concept of semantic similarity is being used in two features to judge both relevance of the essay to the topic and the coherence of the essay itself. It is being calculated by using WordNet[7] and NLTK library. The text is first tokenised into sentences and for each pair of sentences, their semantic similarity is computed using a multi-step process.

Step 1: Term pairs are formed and represented as (i,j) such that i belongs to the first sentence and j belongs to the second sentence. The root of both words (I and j) are compared. A semantic score is assigned to the pair by using the WordNet and NLTK library term similarly function. This process is repeated for all possible pairs for the two sentences.

Step 2: Out of all such scores computed in Step 1, the highest score is taken to be the semantic similarity of the two sentences. We repeat the process (Step 2) for all pairs of sentences in the document.

Step 3: The semantic similarity score of the entire piece of text (either paragraphs or the document as a whole) is computed by taking the average of the semantic similarity scores assigned to each pair of sentences.

Step 4: The average score obtained in the previous step is then multiplied by the log (to the base 2) of the number of sentences. This normalisation is done to ensure that essays with very few sentences (and thus far away from the specified word count) do not receive high scores.

Semantic Similarity of the essay with the topic: this feature evaluates the relevance of the essay to the given topic by computing the semantic similarity of each sentence from the topic and each sentence from the essay. A high score implies that the essay is fairly relevant to the topic.

Semantic Similarity of the essay: this feature evaluated the coherence of the essay itself by computing the semantic similarity of each sentence of the essay to every other sentence of the essay. A high score implies that the essay is fairly coherent.

There is a wide range of supervised learning algorithms. Following are the three classifiers used in our experiments:

kNN k-Nearest Neighbour has been widely used in text classification problems. It is a simple and efficient approach. It is called as a lazy learner as it only stores all the training examples in the learning phase. It does not build a statistical model in the training phase. It does the classification by finding the k-closest training examples and doing a weighted or majority vote for predicting the target class (Tan, 2006).

Linear Regression Linear regression based approaches can be used for text classification (Zhang and Oles, 2001). Linear regression based classifier works by selecting a linear discriminant function and then selecting a threshold value for classification (Zhang and Oles, 2001).

SVM Support Vector Machines are supervised learning models and have been used in several types of text classification problems (Smola and Schölkopf, 2004). SVM classifier works by creating a hyperplane separating the instances (represented as points in a space) of the target classes in an n-dimensional feature space. SVM method is good for both linear and non-linear classification problems (Smola and Schölkopf, 2004).

There is a wide range of feature selection and ranking techniques. We use the following two approaches in our experiments.

[7] `https://wordnet.princeton.edu/`

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3518v1 | CC BY 4.0 Open Access | rec: 9 Jan 2018, publ: 9 Jan 2018

Essay Set	GSR	NTE	NSE	AWC	ASC
Set-1	2-12	1783	589	365.89	22.78
Set-7	0-30	1569	441	168.12	11.65
Set-8	0-60	723	233	609.39	34.86

Table 1. Overview of Experimental Dataset. GSR: Grade Score Range, NTE: Number of Training Essays, NSE: Nmber of Test Essays, AWC: Average Word Count, ASC: Average Sentence Count

RELIEF RELEIF is a widely used feature selection algorithm and is based on taking into account the attribute inter-relationship by computing values such as correlation and covariance (Kira and Rendell, 1992)(Kononenko, 1994)(Robnik-Sikonja and Kononenko, 1997). It is based on the concept of attribute estimation in which a relevance grade is assigned to each of the features and selection is based on a threshold value (Kira and Rendell, 1992)(Kononenko, 1994)(Robnik-Sikonja and Kononenko, 1997)

Correlation-based Feature Subset Selection (CFS) This technique was proposed by Hall et al. (Hall, 1998). CFS computes the importance of a subset of attributes by evaluating the individual predictive ability of each of the attributes along with the degree of redundancy between the attributes (Hall, 1998).

There are 9 features or independent variables for our classification problem. The range and scale of all the independent variables are different and hence we apply techniques to standardize the range of our independent variables. Data normalization and scaling is an important data pre-processing step and is done before applying the classification algorithms (Graf et al., 2003). We rescale the range of 9 features to a scale in the range of 0 to 1.

2.2 Experimental Dataset

In this work, we used publically available Hewlett Foundation's Automated Student Assessment Prize (ASAP) dataset for experimental evaluation. Users can freely sign-up on Kaggle and download this dataset. This dataset has been extensively used in literature for evaluating automatic grading techniques (Dong and Zhang, 2016)(Cummins et al., 2016). ASAP dataset is divided into 8 sets, where each set has a different domain. This ensures variability of domain in dataset. Each set comprises labelled training and testing essay data. All essays have been hand graded by 2 to 3 instructors and based on the combined scores of instructors, a final grade has been assigned. We use a publicly available dataset to enable easier reproducibility and replicability of our results (Stodden, 2012)(Vitek and Kalibera, 2011). We believe that experiments on automated essay scoring should be done on a shared data so that the approaches can be compared and improved by other researchers than the inventors of a particular approach (Stodden, 2012)(Vitek and Kalibera, 2011).

Each essay set has a different grade score range (Set 1: 2-12, Set 2: 1-6, Set 3: 0-3, Set 4: 0-4, Set 5: 0-4, Set 6: 0-4, Set 7: 0-30, Set 8: 0-60). Out of 8 available essay sets, we used 3 essay sets (Set 1, Set 7 and Set 8). We selected the essay sets which have highest grade range for experimental evaluation. This allows a wide distribution of grade levels. Overall, we used a total of 4075 essays for training and 1263 essays for testing. Table 1 presents an overview of the experimental dataset. Table 1 shows that the number of training essays is 1783 for Set 1. The number of training essays for Set 1 is the highest amongst the three sets. Table 1 displays the average word count and sentence count for the essays in the three sets. We observe that the average sentence count varies from 11 to 34 across the three sets in our experimental dataset. Table 1 shows that the number of test essays are sufficient to evaluate the accuracy of the proposed approach. We believe that our dataset is diverse (three different sets) and large (4075) to increase the generalizability of our results.

3 RESULTS

3.1 Feature Distribution

Table 2 shows the descriptive statistics presenting the summary of the 9 features in-terms of the central tendency, dispersion and spread for Set 1. Table 2 shows that the median value for the *Vocabulary* is 0.23 and the Tense is 0.56. The median values shows in Table 2 is the measure of the centrality and can

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3518v1 | CC BY 4.0 Open Access | rec: 9 Jan 2018, publ: 9 Jan 2018

	Attribute	Min	Q1	Median	Q3	Max
1	Vocabulary	0.01	0.18	0.23	0.29	0.61
2	Word Count Limit Ratio	0.00	0.15	0.28	0.43	0.98
3	Semantic Similarity Topic Essay	0.00	0.92	1.02	1.11	1.57
4	Voice	0.72	0.96	0.98	1.00	1.00
5	Semantic Similarity Essay	0.00	1.03	1.16	1.31	2.07
6	Spell Errors	0.00	0.02	0.03	0.04	0.53
7	Tense	0.35	0.50	0.56	0.62	1.00
8	Grammatical Errors	0.00	0.01	0.02	0.03	0.11
9	Long Sentences	0.00	0.33	0.46	0.60	1.00

Table 2. Descriptive Statistics for 9 Features for Set 1

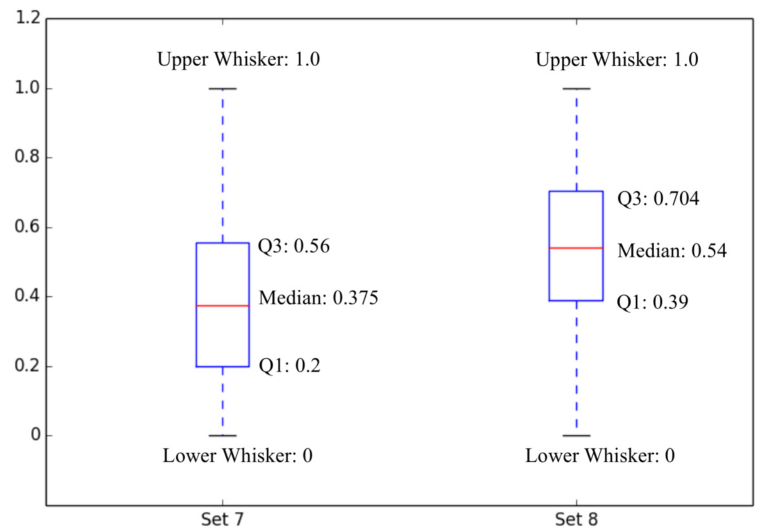

Figure 2. Boxplot for the Attribute Long Sentences for Set 7 and Set 8

provide insights on the skewness of the data. Table 2 displays the first and third quartile values ($Q1$ and $Q3$) which can be used to compute the interquartile range indicating the variability around the mean. We compute the descriptive statistics in Table 2 to observe data patterns and generate hypothesis. We show the descriptive statistics in Table 2 for only one Set (Set 1) as an illustration. We observe variability and spread in the feature values for the other Sets also. Table 2 shows that the $Q1$ value of Voice is 0.96 and the $Q1$ value of the Spell Errors is 0.02. We observe that the $Q3$ value of Semantic Similarity Topic Essay is 1.11. From the numerical summary we infer that the values for the 9 features are scattered and have a spread. The feature values are diverse and contains several values between the largest and the smallest.

Figure 2 and 3 shows the boxplots for displaying and comparing the distribution of two attributes across two datasets. Figure 2 shows a comparison of the attribute Long Sentences for Set 7 and Set 8. Figure 3 shows a comparison of the attribute Tense for Set 7 and Set 8. The boxplot in Figures 2 and 3 presents the five number summary: minimum, first quartile, median, third quartile, and maximum. Figure 2 reveals that the median value of Long Sentences for Set 8 is higher than the medial value for Set 7. Similarly the Q1 and Q3 values of Long Sentences for Set 8 is higher than the Q1 and Q3 values for Set 7. The datasets in Figures 2 and 3 spans the same range since we normalized the values (between 0 and 1). The boxplots in in Figures 2 and 3 in comparing the distributions and shows that there is a variation in the values of a feature across datasets and within a dataset. The boxplot in Figure 3 reveals a very less difference in the middle portion of the feature values across the two Sets. The Tense value of 0.518 divides the dataset into two halves for Set 7 and the tense value of 0.459 divides the dataset into two

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3518v1 | CC BY 4.0 Open Access | rec: 9 Jan 2018, publ: 9 Jan 2018

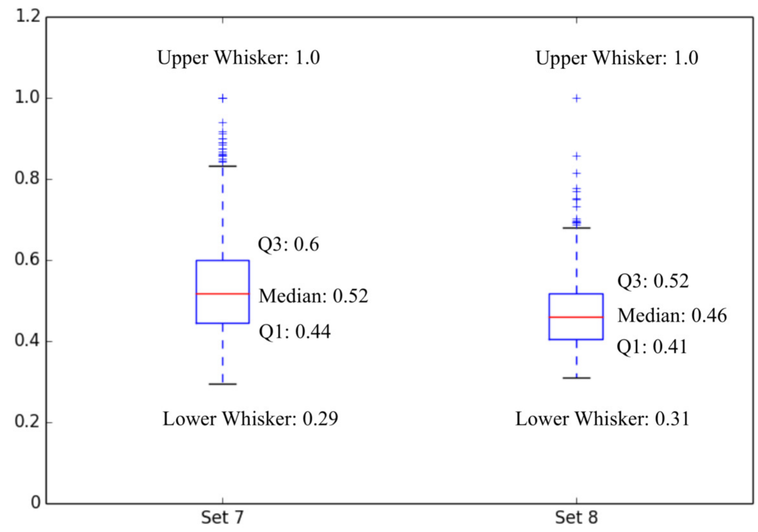

Figure 3. Boxplot for the Attribute Tense for Set 7 and Set 8

halves for Set 8. The interquartile range denotes the middle half of the dataset. The interquartile range and the data distribution of Tense feature shows a similar skewness pattern. We observe that both the boxplots shows a symmetric skewness pattern which is important to understand from the perspective of building predictive models.

3.2 Feature Selection

Item	Quantity	Attribute
1	0.0065584	Vocabulary
2	0.0047464	Word Count Limit Ratio
3	0.0026625	Semantic Similarity Topic Essay
4	0.002551	Voice
5	0.0007761	Semantic Similarity Essay
6	0.0007395	Spell Errors
7	0.0004207	Tense
8	0.0000431	Grammatical Errors
9	-0.0018502	Long Sentences

Table 3. Ranking Results from Relief Feature Selection Algorithm

Feature selection is an important pre-processing step in the machine learning based classification data processing pipeline. We apply feature selection to identify features which are informative and remove redundant or irrelevant features. We use two different features selection techniques (1) RELIEF (2) Correlation-based Feature Subset Selection (CFS). Our objectives behind the application of feature selection technique is to also gain insight about the strength of relationship between the feature and the target class. We apply two different types of feature selection techniques: one of the techniques ranks the features (RELIEF) and the other technique does not rank but identifies a subset of most relevant features (CFS). We use RELIEF feature selection algorithm which can be applied to both binary and continuous data (Kira and Rendell, 1992)(Kononenko, 1994)(Robnik-Sikonja and Kononenko, 1997). RELIEF was proposed by Kira and Rendell et al. (Kira and Rendell, 1992) and then updates to the algorithm was made by Kononenko et al. (Kononenko, 1994). We use the updated version of the RELIEF feature selection algorithm implemented in Weka machine learning software. RELEIF based feature selection techniques are able to detect feature dependencies also. RELEIF algorithm evaluates the importance or worth of the

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3518v1 | CC BY 4.0 Open Access | rec: 9 Jan 2018, publ: 9 Jan 2018

Confusion Matrix for Set 1 (kNN Classifier)

		Predicted Class			
		Class A	Class B	Class C	Class D
Actual Class	Class A	75	66	0	0
	Class B	19	380	13	0
	Class C	0	8	26	0
	Class D	0	0	0	2

Confusion Matrix for Set 7 (SVM Classifier)

		Predicted Class			
		Class A	Class B	Class C	Class D
Actual Class	Class A	10	2	0	0
	Class B	8	237	55	0
	Class C	0	27	100	0
	Class D	0	0	2	0

Confusion Matrix for Set 8 (kNN Classifier)

		Predicted Class			
		Class A	Class B	Class C	Class D
Actual Class	Class A	1	1	0	0
	Class B	9	210	4	0
	Class C	0	2	6	0
	Class D	0	0	0	0

Table 4. Confusion or Error Matrix for the Best Performing Classifier on a Dataset

feature and assigns a weight to it. Table 3 shows the rank of the nine features selected from an initial list of 15 features and their corresponding weights assigned by the RELIEF algorithm. The weights are computed by a process of repeatedly sampling an instance in the dataset and analysing or computing the value of the given feature for the nearest neighbour of either the same or the differential class (Kira and Rendell, 1992)(Kononenko, 1994)(Robnik-Sikonja and Kononenko, 1997). Table 3 reveals that the top 5 attributes are: Vocabulary, Word Count Limit Ratio, Semantic Similarity Topic Essay, Voice and Semantic Similarity Essay.

Correlation-based Feature Subset Selection (CFS) was proposed by Hall et al. (Hall, 1998). CFS computes the importance of a subset of attributes by evaluating the individual predictive ability of each of the attributes along with the degree of redundancy between the attributes (Hall, 1998). According to the CFS technique, subsets of features or attributes in the dataset that are highly correlated with the target class while having low inter-correlation or inter-association are preferred (Hall, 1998). The result of applying CFS was the subset containing four attributes: Long Sentences, Tense, Semantic Similarity Topic Essay and Vocabulary. In our case, ReliefF ranks Vocabulary and Word count limit ratio very high. These two attributes have a good correlation and hence only one of them (Vocabulary) appears in the subset that CFS output. CFS algorithm selects 4 attributes out of which both the algorithms agree on 2 attributes i.e., Vocabulary and Semantic Similarity Topic Essay. This indicates that the two attributes are good predictors of the essay score. The subset produced by CFS contains Long Sentences as well as Tense which are ranked relatively low by RELIEF. This is due to the fact that CFS also checks for low intra-correlation but Relief ranks them individually.

3.3 Confusion or Error Matrix

Table 4 shows the confusion or error matrix displaying the performance of the best performing classifier on a particular dataset. We discretize the score into four categories: A, B, C and D. For example, in-case of Set 1 an ‘A’ grade represents a score of 10-12, ‘B’ grade represents a score of 7-9, ‘C’ grade represents a score of 4-6 and ‘D’ grade represents 2-3. The row represents the actual class and the column represents the predicted class. Table 4 reports the false positives, false negatives, true positives, and true negatives for every category. The confusion matrix is for the test data and shows that the kNN classifier on Set 1

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3518v1 | CC BY 4.0 Open Access | rec: 9 Jan 2018, publ: 9 Jan 2018

Set	Grade	CCI	INC	ACC	OSACC
Set 1	Grade A	75	66	53.19%	82.00%
	Grade B	380	32	92.23%	
	Grade C	26	8	76.47%	
	Grade D	2	0	100%	
Set 7	Grade A	10	2	83.33%	78.68%
	Grade B	237	63	79.00%	
	Grade C	100	27	78.74%	
	Grade D	0	2	0%	
Set 8	Grade A	1	1	50%	93.13%
	Grade B	210	13	94.17%	
	Grade C	6	2	75.00%	
	Grade D	0	0	NA	

Table 5. Accuracy Results for the Best Performing Classifier Across Grade Categories and Sets. CCI – Correctly Classified Instances, INC – Incorrectly Classified Instances, ACC – Accuracy, OSACC – Overall Set Accuracy. Results are for best performing classifier in each set. kNN for Set 1 and Set 8. SVM for Set 7.

Essay Set	kNN	SVM	LR
Set-1	82%	79.62%	80.30%
Set-7	73.69%	78.68%	77.32%
Set-8	93.13%	90.98%	91.41%

Table 6. Overall Classification Accuracy for 3 Classifiers across 3 Dataset

correctly classified 380 instances of the test set actually belonging to the category ‘B’ and misclassified 19 instances into ‘A’ and 13 instances into ‘C’. Table 4 provides a detailed analysis of the quality of the output of the classifier on the essay dataset. The mislabelled or misclassified instances are the off-diagonal elements. For example, for the SVM classifier for Set 7, the number 3, 8, 27 and 57 are off-diagonal elements which are misclassified by the SVM classifier. The diagonal elements represents the number of instances which are correctly classifiers. For example, for the SVM classifier for Set 7, 10, 100 and 237 are correctly classified instances. A higher value of the diagonal elements and a lower value of the off-diagonal elements represents a good classifier. Table 4 reveals very encouraging results for the kNN classifier on Set 8. Table 4 reveals that the number of incorrect predictions for the kNN classifier on Set 8 is very less. For example, in the case of category ‘B’, only 13 instances are misclassified whereas 210 instances are correctly classified.

3.4 Performance Summary and Classifier Comparison

Table 5 shows the detailed performance results for the best performing classifier for the four classes and for the three Sets. Table 5 shows that the accuracy of the kNN classifier for Set 1 with respect to the class ‘B’ is 92.23%. The performance of the kNN classifier for Set 1 is low (53.19%) for class ‘A’ in comparison to the performance of the SVM classifier for for the same class. The performance of the SVM classifier for class ‘A’ is 83.33% for Set 7. The performance of the kNN classifier for class ‘B’ for Set 8 is 94.17%. Table 5 shows that the majority class is ‘B’ and class ‘D’ is a minority class. The best performing classifier for class ‘C’ is SVM with an accuracy of 78.74%. We observe an accuracy of above 75% for the class ‘C’ by all the three best performing classifiers. It is not possible to provide much insight on the classification performance for the class ‘D’ as the number of test instances belonging to class ‘D’ is very less.

Figure 4 displays a histogram to show a visual comparison of the overall accuracy of the three classifiers for the three dataset. The histogram in Figure 4 is derived from information in Table 6. Table 6 and Figure 4 reveals that the overall accuracy for the classifiers kNN, SVM and LR for Set 1 is 82%, 79.62% and 80.30% respectively. kNN is the best performing classifier for Set 1. However, the difference between the performance of the classifier in terms of the overall accuracy is less than 3%. The best

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3518v1 | CC BY 4.0 Open Access | rec: 9 Jan 2018, publ: 9 Jan 2018

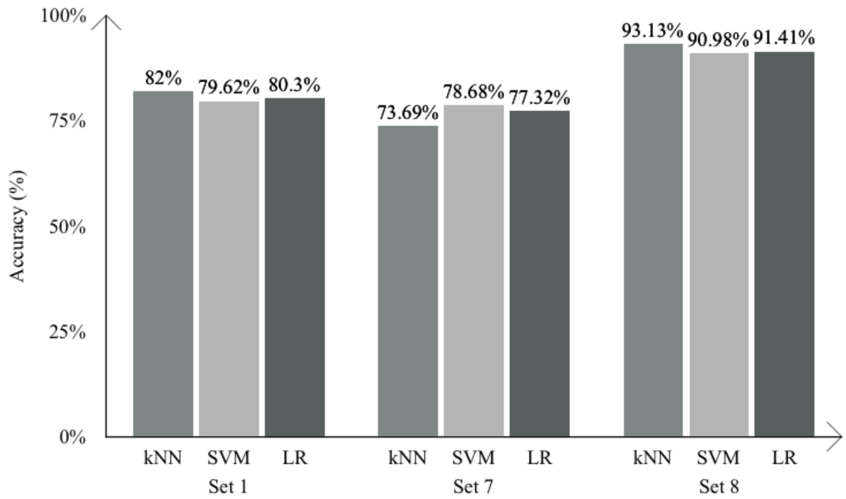

Figure 4. Histogram for Classifier Comparison

performing classifier for Set 7 is SVM with an accuracy of 78.68%. Table 6 and Figure 4 reveals that the overall accuracy for the classifiers kNN, SVM and LR for Set 8 is 93.13%, 90.98% and 91.41% respectively. kNN is the best performing classifier for Set 8. LR is the second best performing classifier for Set 8. We observe different accuracy for the three classifiers but the difference is not substantial and is within 5%. Based on the relative comparison of the three Sets, all classifiers exhibit high accuracy for Set 8 and poor accuracy for Set 7. This result reveals that both the dataset and the classification algorithm influences the accuracy.

4 DISCUSSION

4.1 Interpretation of Findings and Recommendations

Our experimental results shows that it is possible to automatically estimate the writing quality and score of an essay written by school age students using natural language processing and machine learning techniques. The linguistic features as indicators or predictors depends on the essay rubric used by the human judge. For example, if grammar and mechanics (free from spelling, grammar and punctuation errors) is one of the criteria in the grading rubric then then using it as a feature in the machine learning framework will be useful. The score is a function of several features of varying relevance. Few features are more relevant and important than other features and it is interplay or combination of multiple feature values which determines the final score. We observed that both types of features are required: surface level features such as counting words and deep or more sophisticated features such as computing the coherence or writing style. Our insights reveal that it is important to perform feature scaling and normalization as the range and distribution of features vary. We observe that different classifiers result in difference accuracy. Hence more experiments (as future work) is needed to investigate the performance of more classifiers in addition to the three classifiers examined in our study in this paper. The overall accuracy also varies with the dataset. This shows that different features may be needed for different dataset as the grading rubric and level may have some variation across the dataset depending on the context such as the grade level. We observe some imbalance in the dataset with respect to the grade (A, B, C and D). There are very few essays with grade D and majority are in B and C. A and D grade is a minority class. In future data sampling techniques such as oversampling, under=sampling and SMOTE can be applied to counter the class imbalance problem to further enhance the performance of the essay grading the system. The class imbalance is natural as the grade often follows a normal distribution. Our results are encouraging and positive however more research is needed to investigate misclassified and incorrectly classified results. We believe that few misclassification can be corrected by making improvements to the system but few misclassifications are not due to the shortcoming of the automated essay grading system and rather due to subjectivity in evaluation and possible human errors.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3518v1 | CC BY 4.0 Open Access | rec: 9 Jan 2018, publ: 9 Jan 2018

4.2 Threats to Validity

There are several possible threats to validity in our experiments which we tried to minimize and mitigate (internal and external validity threats) (Winter, 2000). We conduct experiments on multiple and diverse dataset belonging to three different projects to investigate if our results are generalizable and hence mitigate the threat to external validity. We downloaded a essay dataset from Kaggle repository which is manually validated and of high quality. The dataset has been used in several experiments in the past and our dataset selection is keeping mind that there are no annotation or measurement errors. However, there is still a possibility of threats to internal validity in such empirical experiments. The impact on the dependent variable (target clas) may not be completely attributed to the changes in the independent variable (input features) because of overfitting of the predictive model. Another threat to validity is that our 9 textual features may not be the only factor leading to writing skill or essay quality and there can be other factors not included as part of our study presented in this paper.

5 CONCLUSIONS

We present machine learning and natural language processing based approach for automated essay grading. We propose 9 surface level and deep linguistic features as predictors for the writing skill of the author and quality of the essay in the context of Grade 7 to Grade 10. We conduct experiments on three sets of publicly available and manually annotated dataset consists of more than 4000 essays across diverse topics. We observe variability and spread in the feature values of the 9 attributes across 3 sets. We apply two different types of feature ranking techniques. We conclude that features such as appropriate use of vocabulary, word count with respect to the world limit, and relevance of the essay to the topic, coherency in writing and correct usage of active and passive voice are good predictors of the essay score. We apply there different classification algorithms to build predictive modes: k-nearest neighbour, support vector machines and linear regression based classifier. Our analysis shows that the accuracy of the kNN classifier for Set 1 with respect to the class 'B' is 92.23%. kNN is the best performing classifier for Set 1. The best performing classifier for Set 7 is SVM with an accuracy of 78.68%. We observe different accuracy for the three classifiers but the difference is not substantial and is within 5%. Our results on 4075 essays across multiple topics and grade score range are encouraging with an accuracy of 73% to 93%.

REFERENCES

Balfour, S. P. (2013). Assessing writing in moocs: Automated essay scoring and calibrated peer review (tm). *Research & Practice in Assessment*, 8.

Bin, L., Jun, L., Jian-Min, Y., and Qiao-Ming, Z. (2008). Automated essay scoring using the knn algorithm. In *Computer Science and Software Engineering, 2008 International Conference on*, volume 1, pages 735–738. IEEE.

Chen, Y.-Y., Liu, C.-L., Lee, C.-H., Chang, T.-H., et al. (2010). An unsupervised automated essay scoring system. *IEEE Intelligent systems*, 25(5):61–67.

Cummins, R., Zhang, M., and Briscoe, T. (2016). Constrained multi-task learning for automated essay scoring. Association for Computational Linguistics.

Dong, F. and Zhang, Y. (2016). Automatic features for essay scoring-an empirical study. In *EMNLP*, pages 1072–1077.

Graf, A. B., Smola, A. J., and Borer, S. (2003). Classification in a normalized feature space using support vector machines. *IEEE Transactions on Neural Networks*, 14(3):597–605.

Ha, M. and Nehm, R. H. (2016). The impact of misspelled words on automated computer scoring: A case study of scientific explanations. *Journal of Science Education and Technology*, 25(3):358–374.

Hall, M. A. (1998). *Correlation-based Feature Subset Selection for Machine Learning*. PhD thesis, University of Waikato, Hamilton, New Zealand.

Kakkonen, T., Myller, N., Timonen, J., and Sutinen, E. (2005). Automatic essay grading with probabilistic latent semantic analysis. In *Proceedings of the second workshop on Building Educational Applications Using NLP*, pages 29–36. Association for Computational Linguistics.

Kira, K. and Rendell, L. A. (1992). A practical approach to feature selection. In Sleeman, D. H. and Edwards, P., editors, *Ninth International Workshop on Machine Learning*, pages 249–256. Morgan Kaufmann.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3518v1 | CC BY 4.0 Open Access | rec: 9 Jan 2018, publ: 9 Jan 2018

Kononenko, I. (1994). Estimating attributes: Analysis and extensions of relief. In Bergadano, F. and Raedt, L. D., editors, *European Conference on Machine Learning*, pages 171–182. Springer.

McNamara, D. S., Crossley, S. A., Roscoe, R. D., Allen, L. K., and Dai, J. (2015). A hierarchical classification approach to automated essay scoring. *Assessing Writing*, 23:35–59.

Nehm, R. H., Ha, M., and Mayfield, E. (2012). Transforming biology assessment with machine learning: automated scoring of written evolutionary explanations. *Journal of Science Education and Technology*, 21(1):183–196.

Robnik-Sikonja, M. and Kononenko, I. (1997). An adaptation of relief for attribute estimation in regression. In Fisher, D. H., editor, *Fourteenth International Conference on Machine Learning*, pages 296–304. Morgan Kaufmann.

Rudner, L. M., Garcia, V., and Welch, C. (2006). An evaluation of intellimetric™ essay scoring system. *The Journal of Technology, Learning and Assessment*, 4(4).

Smola, A. J. and Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and computing*, 14(3):199–222.

Stodden, V. (2012). Reproducible research for scientific computing: Tools and strategies for changing the culture. *Computing in Science & Engineering*, 14(4):13–17.

Tan, S. (2006). An effective refinement strategy for knn text classifier. *Expert Systems with Applications*, 30(2):290–298.

Vitek, J. and Kalibera, T. (2011). Repeatability, reproducibility, and rigor in systems research. In *Proceedings of the ninth ACM international conference on Embedded software*, pages 33–38. ACM.

Winter, G. (2000). A comparative discussion of the notion of'validity'in qualitative and quantitative research. *The qualitative report*, 4(3):1–14.

Yannakoudakis, H., Briscoe, T., and Medlock, B. (2011). A new dataset and method for automatically grading esol texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 180–189. Association for Computational Linguistics.

Zhang, T. and Oles, F. J. (2001). Text categorization based on regularized linear classification methods. *Information retrieval*, 4(1):5–31.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3518v1 | CC BY 4.0 Open Access | rec: 9 Jan 2018, publ: 9 Jan 2018