A Graph Based Approach to Automate Essay Evaluation

Reecha Bhatt*, Malvik Patel *, Gautam Srivastava†, Vijay Mago*
*Department of Computer Science, Lakehead University, Canada
†Department of Mathematics and Computer Science, Brandon University, Manitoba, Canada

***Abstract*—Despite studies of over six decades, the research on automated essay scoring continues to grab ample attention in the natural language processing (NLP) community in part because of its commercial and educational values.However, evaluating such writing compositions or essays in terms of reliability and time is a very challenging process. The need for accurate and rapid scores has elevated the need for a computer program that can automatically evaluate essay questions that match particular prompts. Automated Essay Scoring (AES) systems are used by the NLP and machine learning strategies to solve the difficulties of scoring writing tasks. In this paper, we suggest an AES approach that involves not only rule-based grammar and consistency tests, but also the semantic similarity of sentences, thus giving priority to question prompts. We have used similarity vectors obtained after applying semantic algorithms and calculated statistical features. Our system uses 22 features with high predicting power, which is less than current systems, while considering every aspect a human grader may focus on.Predicting scores is achieved using the data provided by Kaggle's ASAP competition using Random Forest. The resulting agreement between the score of the human grader and the prediction of the system is compared with promising results through experimental evaluation.**

***Index Terms*—NLP, semantics, statistics, syntax, essays, evaluation, automated**

I. Introduction

There are multiple modes for assessing student learning effectively, but writing comes at the top, especially in academic contexts. Consequently, In many academic disciplines the writing of essays is used as a method of evaluation because it is easier to assess student learning using writing tasks [1]. However, writing prompts also take longer to evaluate compared to other assessment methods such as multiple choice tests. The importance of automated essay scoring is evident through the increased need for education in our well educated workforces [2]. Furthermore, the subjective nature of current essay assessment often results in unfariness towards the student [3]. Also, it is a hectic and boring job for teachers. Therefore, using Automated Essay Scoring (AES) systems is important as it increases our understanding of textual features and cognitive abilities involved in collecting and interpreting written texts, providing several benefits to the educational community. More importantly students can evaluate themselves anytime they want to; human graders may not be available all the time to provide feedback and accurate scores. As a result, students can save money which they typically give to institutions for evaluating their progress before final exams. Overall, AES systems provide multiple benefits to teachers, educational institutions, and students. Current AES systems are plagued with major gaps between their results compared to human grading. This paper looks to bridge that gap.

In this paper, we present a novel AES system, through which human grading may be replaced with machine grading. We propose an automated essay grading system with a smaller number of features but with high predicting power omitting feature with less predicting power to eliminate redundancies and improve accuracy. We have extracted the features based on previous studies. Features were selected based on crucial characteristics of a good essay. As seen in Fig. 1, the main features extracted are based on statistical, semantic and syntactic analysis. These features were tested on different supervised prediction models to find out which model works the best.

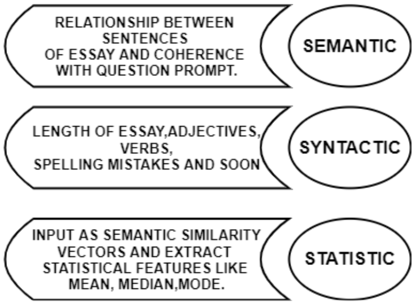

RELATIONSHIP BETWEEN SENTENCES OF ESSAY AND COHERENCE WITH QUESTION PROMPT.
SEMANTIC
LENGTH OF ESSAY,ADJECTIVES, VERBS, SPELLING MISTAKES AND SOON
SYNTACTIC
INPUT AS SEMANTIC SIMILARITY VECTORS AND EXTRACT STATISTICAL FEATURES LIKE MEAN, MEDIAN,MODE.
STATISTIC

Figure 1: Extracted features

This paper also addresses the latest methods in the automated assessment of English essays and the successful use of these frameworks as a basis for the new system. The well known systems for AES which are all available commercially are as follows:

1) C-Rater [4]
2) Intelligent Essay Marking System [5]
3) Intelligent Essay Assessor [6]

4) Educational Testing service
5) Electronic Essay Rater (ERater)
6) Project Essay Grade (PEG) [7]

II. Related Work

Page showed as early as 1966 that an automated ”rater” is comparable with human raters [8]. In 1982, a UNIX programmer named Writer’s Workbench may provide guidance on punctuation, pronunciation, and grammar. In the 1990s more systems were created, with Intelligent Essay Assessor the most prominent systems [9], Intellimetric [10], a new version of the Project Essay Grade [8], and e-rater [11]. Page set the stage for automated writing evaluation with an automated essay grading system called Project Essay Grader (PEG) [8]. It was first used commercially in 1999.

In 2015, McNamara *et al.* used a hierarchical classification approach in which they also added feedback for students [12]. In 2016, a neural approach was proposed by Taghipour *et al.* for automated essay scoring [13]. In 2017, Yamamoto presented an interesting study on AES systems based on a rubric using five evaluation viewpoints which were: Content, Structure, Evidence, Style, and Skill [14]. We have also seen strong recent work by Janda *et al.* in [15] which we directly build on here.

A. Semantic Analysis

In the point of view of data processing, semantics can be defined as ”tokens’ that are able to provide language context. Semantics are able to offer clues to word meanings as well as their relationships with other words. It is the process of relating syntactical structure, from phrases, clauses and sentences to the level of writing. For essay evaluation semantic analysis is used to find coherence among sentences of essays and between the question prompt and the whole essay which is very effective to grade essay writing skills. There have been different approaches in extracting semantic information using several NLP techniques. Supervised and Unsupervised learning approaches are used in [16]. Some systems have used Latent Semantic Analysis (LSA) to determine coherence of texts which is fully automated [17], [18].

B. Statistical Analysis

The Intelligent Essay Assessor [19] is based on Latent Semantic Analysis [20], a statistical technique for summarizing the relations between words in documents. In the first version of e-rater in 1998, they used a stepwise regression technique to select the best features that are most predictive for a given set of data. PEG is also based on regression analysis [21].

After considering shortcomings of statistical analysis on AES we have tried a new approach which will be discussed further in this paper.

III. Methodology

A. Study Design

Feature extraction methods utilizing machine learning algorithms was preferred as the feature selection and extraction methods are useful for different applications such as face recognition, action, recognition, gesture recognition, biomedical engineering, marketing, and wireless network [22]. The variety of applications of feature selection and extraction have shown their usefulness and effectiveness in different real-world problems. Moreover, machine learning, by nature, is a mathematical approach to problem solving, and it makes heavy use of statistical and discrete analyses relying on numerical conversions and probability theory [23]. Use of quantitative analyses have several benefits: it allows broader study, involving a greater number of subjects, and enhancing the generalisation results. It also allows for greater objectivity and accuracy of results. Thus, Considering the advantages of the quantitative research methods and the nature of machine learning, the research problem was approached from the quantitative perspective in the present study. The following subsections have the step wise procedure with the flow of the system.

B. Procedures

The methodology in this paper has the following steps, also summarized in Figure 2.

1) *Read Dataset:* In this step, the dataset being used in the system is read.
2) *Preprocessing of the data:* After the dataset is selected, the text is processed to remove extra whitespaces, convert accented characters to ASCII characters, expand contractions, remove special characters, change the case of the text to lowercase.
3) *Data cleaning:* The cleaning process includes tokenization, capitalization or de-capitalization, removing stop words, breaking attached words, lemmatization/stemming. It also includes spell check and grammar correction but we ignored it for this project as to better evaluate the mistakes in the essays.
4) *Feature extraction:* Extracting features with high predicting powers will lead to overall better performance and accuracy of the model hence we have tried extracting a lot of features from different domains which will be explained further in the research paper.
5) *Feature selection:* After feature extraction, the selection of more predictive features is of vital importance. There are many techniques which can be used, like univariate selection, information gain and correlation

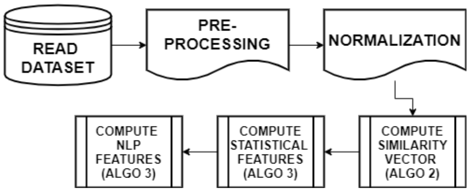

READ DATASET
PRE-PROCESSING
NORMALIZATION
COMPUTE SIMILARITY VECTOR (ALGO 2)
COMPUTE STATISTICAL FEATURES (ALGO 3)
COMPUTE NLP FEATURES (ALGO 3)

Figure 2: Flow of the System

matrix with heatmap. We have performed two of them which are discussed further.

6) *Testing on a prediction model:* We have trained and tested our data on five supervised machine learning models which are random forest, KNN, SVM and decision tree. We have also tried different attributes and thousands of models with different parameters to find the best accuracy.
7) *Checking accuracy (kappa score):* This is the last step where QWK scores are evaluated based on the similarity between the predicted values and the ground truth.

C. Feature Extraction Process Flow

After reading the dataset, the first step is cleaning the dataset by removing the unwanted punctuation, tags and stop words ("the", "is"). It also includes stemming, lemmatization, and normalizing the target variable and bring it in the range of 0 to 1 then applying algorithms to extract features. The use of feature extraction is better over other methods as it helps to reduce the number of resources needed for processing without losing important or relevant information which reduces the amount of redundant data for given analysis, as a result it makes the model more efficient and feature selection helps in reducing the use of overall computational power.

D. Syntactical Attributes

Syntax refers to the combination of phrases, clauses and sentences, and the different parts of speech that are nouns, verbs and prepositions. Natural processing tool kit (NLTK) a python-based library to extract syntax related features [24] such as tokenization, stemming, lemmatization, punctuation, character count, and word count. This process is summarized in Fig. 3. The process for finding syntactical attributes: first the text is separated into sentences by full stop. Then the sentence is tokenized into words and we count the number words. Finding syntactical attributes is summarized with Algorithm 1, and are listed below:

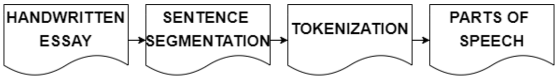


Figure 3: Syntactic flow

Algorithm 1 Construct NLP features

```
Input: Plain Text
Output: Data Frame
procedure GET_NLP_FEATURES
    while sentence in essay do
        D[w_list] ← tokenize(sentences)    ▷ extracting word from essay
        D[l] ← length(w_list)    ▷ calculating length of word list
        D[m_words] ← length(misspelled_words)    ▷ calculating misseplled words
        tagged ← pos(sentence)  ▷  part of speech tagging
        D[u_pos] ← len(tagged)    ▷ unique pos tags
        D[ex] ← tagged(ex)    ▷ extract existential there
        D[jj] ← tagged(jj)    ▷ extract adjective
        D[pdt] ← tagged(pdt)    ▷ extract predeterminer
        D[cc] ← tagged(cc)    ▷ extract coordinating conjunction
        D[jjs] ← tagged(jjs) ▷ extract superlative adjective
        D[nn] ← tagged(nn)    ▷ extract nouns
        D[jjr] ← tagged(jjr)    ▷ extract comparative adj.
        D[vb] ← tagged(vb)    ▷ extract verbs
        D[c_len] ← length(essay)  ▷ calculating length of essay
    return D
```

- *Unique parts-of-speech:* Overuse or repetition of parts of speech is regarded as poor grammar skills. Here the essay is tokenized into words using NLTK. Each word is tagged, identified words are placed in sets, and there is no duplication.
- *Mispplled words:* The use of incorrect spellings can result in the essay evaluator misinterpreting the word. On the spell check library pyEn chant a dictionary for American English is used to find misspelled words. We count the total occurrences of mispelled words.
- *Adjective:* Adjectives are basic elements of sentences in any language. Using adjectives means we can convey the quality of any person or thing. The total number of occurrences of adjectives by JJ tag in the part of the speech library.
- *Character count:* It is also important to test the number of characters, as it shows the overall use of alphabets and non-alphabet components.
- *Predeterminants:* The use of predeterminants is considered as high-quality writing thus we have included this feature.
- *Coordinating Conjunctions (CC):* These are used to incorporate two principal clauses. Larger sentences are harder to comprehend which in turn leads to lower scores for essays. Total number of CC used in the essay are counted.
- *Superlative adjectives (SA):* Total count of SA occurrences is kept and tagged.
- *Ending with -ing:* Excess use of "-ing" makes writing look poor. Total count of this is kept.

- *Verbs:* Verbs are a very important part of speech because without them a sentence cannot exist.
- *Noun:* Nouns are important because they refer to places, objects and people and the more sophisticated abstract concepts. Without nouns, the sentence will be left with verbs, adjectives and adverbs.
- *Number of words:* This count is kept as well and applied in scenarios where a word limit was imposed by a teacher.

E. Semantic Attributes

"Semantic" the word means *meaning or logic*. Text semantic matching is commonly used in many applications, such as computer translation, automated answering of questions, and information recovery. Semantic similarity is useful when one combines related terms into semantic concepts with the same meaning [25]. Similarly, in this research the essay not only should be grammatically correct but also it needs to be related to the asked question (question prompt), and it is the reason why considering semantic features becomes important. It is important that all the parts of the essay fit together to form a whole meaningful thought; therefore, we decided to find similarities between all the sentences of the essay. In addition, as the coherence of the essay with the question prompt is of equal importance, we compared each sentence with the prompt; then after we selected the one sentence with the most similarity and again compared that sentence with each of the other sentences then generated the matrix of similarity vectors. Figure 4 shows the flow and calculation of the similarity vectors where $S_1, S_2, ..., S_n$ are the sentences and $4, 5, 7, ...$ are the similarity scores. With the help of this similarity vector we can drive the novel statistical features. We retrieve the most influencing sentences using Algorithm 2.

Algorithm 2 Get most influencing sentence

```
    Input: Essay
    Output: 1D Array Vector
procedure GET_SIMILARITY
    sentences ← tokenise(essay)                          ▷ seprate sentences
    n ← length(sentences)                                ▷ number of sentence
    x ← [...]                                            ▷ array to store similarities
    prompt_array ← question prompts
    while sentence in essay do
        x ← similarity(prompt, sentence)                 ▷ similarity between prompt and sent.
    max_similarity ← max(similarity_vector)              ▷ max similarity
    y ← maximum similarity sentence                      ▷ sentence with max similarity
    while sentence in essay do
        x ← similarity(y, sentence)                      ▷ similarity between selected sent. and sent.
    return x
```

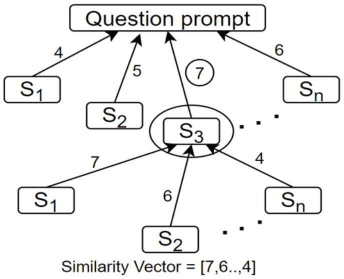


Figure 4: Similarity flow

F. Statistical Attributes

In this study with the help of similarity vectors we have extracted features with high predicting powers all statistical attributes. These attritibutes have been shown to be effective in other AES systems. To construct statistical feature we use Algorithm 3, and are listed below:

Algorithm 3 Construct Statistical features

```
    Input: 1D Array Vector
    Output: Data Frame
procedure GET_STATISTICAL_FEATURES
    D[std] ← stdev(similarity_vector)
    D[mean] ← mean(similarity_vector)
    D[median] ← median(similarity_vector)
    D[mode] ← mode(similarity_vector)
    D[h_mean] ← hmean(similarity_vector)
    D[g_mean] ← gmean(similarity_vector)
    D[variance] ← variation(similarity_vector)
    D[skew] ← skew(similarity_vector)
    D[iqr] ← iqr(similarity_vector)
    if length_of_similarity_vector ≥ $\theta$ then       ▷ where $\theta = 5$
        D[z_score] ← 0
        D[p_score] ← 0
    else
        D[z_score] ← kurtosistest(similarity_vector)[1]
        D[p_score] ← kurtosistest(similarity_vector)[0]
    return D
```

- *Variance:* Variance is a discrepancy between two sentences or text which we believe is important as it shows how much it conflicts with what is asked in the prompt and the essay.
- *Skew:* Skewness lets you test by how much the overall shape of a distribution deviates from the shape of the normal distribution.
- *Z-SCORE:* A z-score (also called a standard score) gives an idea of how far from the mean a data point is. But more technically, it is a measure of how many standard deviations below or above the population mean a raw score is.
- *Geometric Mean:* In statistics, the geometric mean is determined by elevating the sum of a sequence

of the numbers to the inverse of the sequence total length.

- *Harmonic Mean:* One type of numerical average is the harmonic mean. It is calculated by dividing the number of observations of each number in the series by the reciprocity.
- *Average:* A number that expresses the central or typical value in a collection of data, in particular the mode, average or most commonly determined by dividing the sum of the values in the collection by their number.
- *Median:* The median is a pure central tendency metric. We organize the observations to find the median in order from the smallest to the greatest value.
- *Mode:* The most occurring number or value in the population or the sample. we can relate it to the repetitive words which is brings down the scores.
- *IQR:* It is defined as a difference between the largest and the smallest values in the middle of 50 percent of a set data.
- *Standard Deviation:* The standard deviation is a numerical value used to indicate how widely individuals in a group vary. If individual observations vary greatly from the group mean, the standard deviation is big; and vice versa.

IV. Experimental Results

A. Dataset

We use data from the automated student evaluation to verify and compare our findings to the existing systems[1]. The dataset consists of 8 different argumentative (Arg), sensitive (Res), narrative (Nar), persuasive (Per) and expository (Exp) data sets in various genres. Each data set is a selection of their own prompt responses. Students from Grades 6 to 10 wrote essays ranging from 150 to 550 words per essay. A total of 2 graders were given the grades from exams that had:

- essayid: A unique identifier for each individual student essay
- essayset: 1-8, an id for each set of essays as shown in Table I
- Essay: The ASCII text of a student's response
- Domain (1,3-8) prediction-id: A unique prediction-id that corresponds to the predicted-score on the essay for domain 1 (all essays have this)
- Domain2predictionid: A unique predictionid that corresponds to the predictedscore on the essay for domain 2 (only SET2 essays have this)

Kaggle dataset has the following properties.

- Training set size is around 12,978 handwritten essays

Dataset	Shape	Size(kb)
SET1	(1783,5)	263
SET2	(1800,5)	196
SET3	(1726,5)	195
SET4	(1772,5)	195
SET5	(1805,5)	195
SET6	(1805,5)	195
SET7	(1569,5)	196
SET8	(0723,5)	123

Table I: Dimension of the dataset

- Dataset is divided into eight separate datasets
- 8 separate datasets have 8 different question prompts
- 8 of them have different dimensions

B. Feature Selection

Feature selection [31] helps in reducing the model complexity and makes is faster for complex machine learning tasks. It increases a model's accuracy when the appropriate subset is picked. There are several benefits of feature selection such as reduced overfitting, improved accuracy, and reduced training time. There are many methods with which feature selection can be performed. Here we have used the following methods:

(1) Correlation matrix with heat map
(2) Univariate selection

1) Correlation Matrix with Heat Map: Correlation matrix is how features are correlated with the target variable and heatmap makes it easier to identify the most correlated features with its colourful presentation [32]. Figure 5 shows our heatmap where green indicates the most correlation with target variable score whereas red indicates less correlation.

2) Univariate Selection: In this test, we select those features which have the highest correlation with the output variable, and we get the ranking of the features according to the features that are eliminated or kept in the data set. The features with less correlation are deleted as it also cleans the data by removing the redundancies and as a result accuracy improves. Figure 6 shows the correlation. Here the "score" is the output variable.

C. Evaluation Metric

In this study, extensive evaluation was conducted to obtain high accuracy, and the model giving the best results were different in case of time analysis and in case of (QWK) kappa scores. Quadratic Weighted kappa (QWK) is the evaluation metric used to evaluate our method, as this was the official evaluation metric chosen by the ASAP competition. QWK is a calculation on which two

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

System	No. of Features	D-1	D-2	D-3	D-4	D-5	D-6	D-7	D-8	Avg.
New Approach	22	0.82	0.73	0.71	0.72	0.78	0.72	0.70	0.65	0.73
Janda *et al.* [15]	23	0.83	0.81	0.77	0.72	0.87	0.82	0.75	0.78	0.79
SBLSTMA [26]	14 plus several sub-features	0.86	0.73	0.78	0.82	0.84	0.82	0.81	0.75	0.80
SVMrank [27]	33	0.80	0.68	0.67	0.73	0.80	0.71	0.77	0.71	0.73
SKIPFLOW [28]	14 plus several sub-features	0.83	0.68	0.69	0.79	0.81	0.810	0.800	0.69	0.76
Tpioc-BiLSTM-attention[86]	Not mentioned	0.82	0.69	0.69	0.81	0.81	0.82	0.80	0.70	0.77
e-rater [11]	46	0.82	0.69	0.72	0.80	0.81	0.75	0.81	0.70	0.77
IntelliMetric [10]	400	0.78	0.68	0.73	0.79	0.83	0.76	0.81	0.68	0.76
BLRR [29]	15 plus several sub-features	0.76	0.60	0.62	0.74	0.78	0.77	0.73	0.62	0.70
TDNN [30]	Not clear	0.76	0.68	0.62	0.75	0.73	0.67	0.65	0.57	0.68

Table II: Comparing results with related work

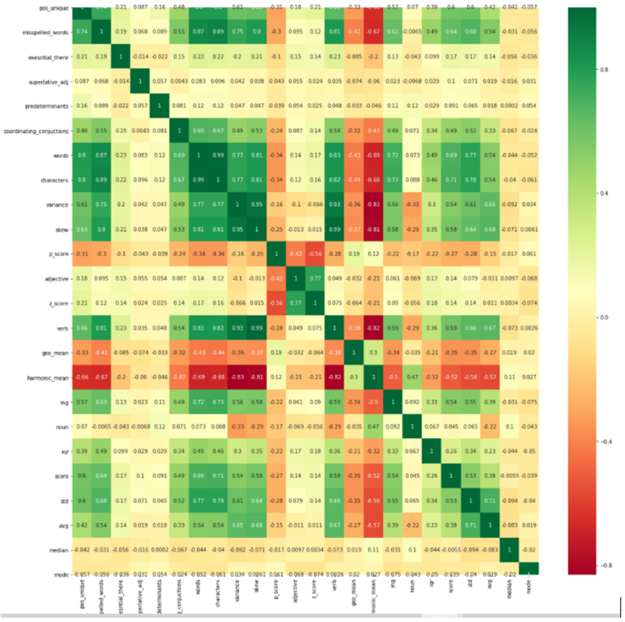
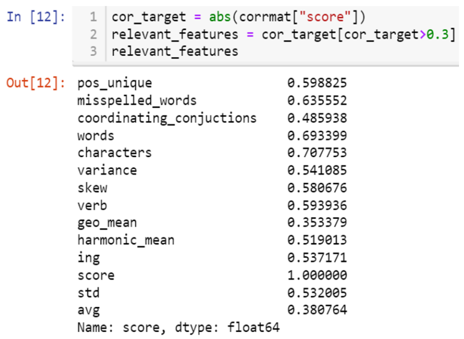


Figure 5: Heatmap



```
In [12]:  cor_target = abs(cormat["score"])
          relevant_features = cor_target[cor_target>0.3]
          relevant_features

Out[12]: pos_unique                  0.598825
         misspelled_words            0.635552
         coordinating_conjuctions    0.485938
         words                       0.693399
         characters                  0.707753
         variance                    0.541085
         skew                        0.580676
         verb                        0.593936
         geo_mean                    0.353379
         harmonic_mean               0.519013
         ing                         0.537171
         score                       1.000000
         std                         0.532005
         avg                         0.380764
         Name: score, dtype: float64
```

Figure 6: Ranking

DATASET	KNN	RF	SVM	DT
SET1	2333	1502	2131	0980
SET2	1149	1726	1016	0865
SET3	1364	2329	0953	0794
SET4	2033	1453	0910	0780
SET5	1535	1551	0961	0852
SET6	1619	1438	0998	1143
SET7	1682	2112	1038	2345
SET8	1278	0981	1590	3410
AVG	1625	1448	1199	1396

Table III: Time Analysis of each model

raters agree. In case of an essay evaluation program, the agreement between the system's expected score and the human rater rating is the QWK, which varies from values 0 (no agreement) to 1 (full agreement).

D. *Time Analysis*

Time analysis is another aspect of machine learning if one model has high accuracy but takes a lot of time whereas the other model has less accuracy but takes lesser time then one may choose the later model. Table III presents the time analysis on the data in milliseconds.

When compared with the related work on the dataset used, we are slightly lower with a 73% (QWK) score with 22 features, see Table II. As can be seen from the results, when compared to previous work we are slightly behind in accuracy however our results were achieved using the least number of features and with the most efficient runtime. We achieve least accuracy in SET8 whereas best accuracy for SET1. Moreover, we were successful to find the highest predicting power features as shown in Fig. 6.

V. Conclusion

In this paper, we present a method that not only integrates rule-based grammar and syntax tests effectively, but also the semantic similarities within the essay demonstrating its coherence specially with the essay prompt. We suggest using statistical features that are based on relationships within the context of the essay. To remove redundancy, we eliminated less correlated features. Our work can be used to replace the manual grading system and free teachers to do other interesting jobs. For future work, sentiment analysis can be added to the model as it is currently missing. The joining of all features with sentiment analysis may in turn produce better results down the road.

References

[1] K. Cho, C. D. Schunn, and R. W. Wilson, "Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives." *Journal of Educational Psychology*, vol. 98, no. 4, p. 891, 2006.

[2] D. Higgins, J. Burstein, D. Marcu, and C. Gentile, "Evaluating multiple aspects of coherence in student essays," in *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, 2004, pp. 185–192.

[3] M. D. Shermis and J. Burstein, "Using automated scoring to monitor reader performance and detect reader drift in essay scoring," in *Handbook of Automated Essay Evaluation*. Routledge, 2013, pp. 255–272.

[4] C. Leacock and M. Chodorow, "C-rater: Automated scoring of short-answer questions," *Computers and the Humanities*, vol. 37, no. 4, pp. 389–405, 2003.

[5] S. Valenti, F. Neri, and A. Cucchiarelli, "An overview of current research on automated essay grading," *Journal of Information Technology Education: Research*, vol. 2, no. 1, pp. 319–330, 2003.

[6] P. W. Foltz, D. Laham, and T. K. Landauer, "The intelligent essay assessor: Applications to educational technology," *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*, vol. 1, no. 2, pp. 939–944, 1999.

[7] E. B. Page, "The use of the computer in analyzing student essays," *International review of education*, pp. 210–225, 1968.

[8] ——, "The use of the computer in analyzing student essays," *International review of education*, pp. 210–225, 1968.

[9] P. W. Foltz, W. Kintsch, and T. K. Landauer, "The measurement of textual coherence with latent semantic analysis," *Discourse processes*, vol. 25, no. 2-3, pp. 285–307, 1998.

[10] S. Elliot, "Intellimetric: From here to validity," *Automated essay scoring: A cross-disciplinary perspective*, pp. 71–86, 2003.

[11] S. M. Lottridge, E. M. Schulz, H. C. Mitzel, M. Shermis, and J. Burstein, "Using automated scoring to monitor reader performance and detect reader drift in essay scoring," *Handbook of Automated Essay Evaluation: Current Applications and New Directions, MD Shermis and J. Burstein, Eds. New York: Routledge*, pp. 233–250, 2013.

[12] D. S. McNamara, S. A. Crossley, R. D. Roscoe, L. K. Allen, and J. Dai, "A hierarchical classification approach to automated essay scoring," *Assessing Writing*, vol. 23, pp. 35–59, 2015.

[13] K. Taghipour and H. T. Ng, "A neural approach to automated essay scoring," in *Proceedings of the 2016 conference on empirical methods in natural language processing*, 2016, pp. 1882–1891.

[14] M. Yamamoto, N. Umemura, and H. Kawano, "Proposal of japanese vocabulary difficulty level dictionaries for automated essay scoring support system using rubric," *Journal of the Operations Research Society of China*, pp. 1–17, 2019.

[15] H. K. Janda, A. Pawar, S. Du, and V. Mago, "Syntactic, semantic and sentiment analysis: The joint effect on automated essay evaluation," *IEEE Access*, vol. 7, pp. 108 486–108 503, 2019.

[16] M. Gamon, "Graph-based text representation for novelty detection," in *Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing*. Association for Computational Linguistics, 2006, pp. 17–24.

[17] M. N. Jones, J. Willits, S. Dennis, and M. Jones, "Models of semantic memory," *Oxford handbook of mathematical and computational psychology*, pp. 232–254, 2015.

[18] S. M. Lottridge, E. M. Schulz, H. C. Mitzel, M. Shermis, and J. Burstein, "Using automated scoring to monitor reader performance and detect reader drift in essay scoring," *Handbook of Automated Essay Evaluation: Current Applications and New Directions, MD Shermis and J. Burstein, Eds. New York: Routledge*, pp. 233–250, 2013.

[19] P. W. Foltz, L. A. Streeter, and K. E. Lochbaum, "Implementation and applications of the intelligent essay assessor," in *Handbook of automated essay evaluation*. Routledge, 2013, pp. 90–110.

[20] V. K. Gupta, P. J. Giabbanelli, and A. A. Tawfik, "An online environment to compare students' and expert solutions to ill-structured problems," in *International Conference on Learning and Collaboration Technologies*. Springer, 2018, pp. 286–307.

[21] P. W. Foltz, L. A. Streeter, and K. E. Lochbaum, "Implementation and applications of the intelligent essay assessor," in *Handbook of automated essay evaluation*. Routledge, 2013, pp. 90–110.

[22] Y. Mingqiang, K. Kidiyo, and R. Joseph, "A survey of shape feature extraction techniques," *Pattern recognition*, vol. 15, no. 7, pp. 43–90, 2008.

[23] B. S. Weir *et al.*, *Genetic data analysis. Methods for discrete population genetic data.* Sinauer Associates, Inc. Publishers, 1990.

[24] E. Loper and S. Bird, "Nltk: the natural language toolkit," *arXiv preprint cs/0205028*, 2002.

[25] M. Litvak and M. Last, "Graph-based keyword extraction for single-document summarization," in *Proceedings of the workshop on Multi-source Multilingual Information Extraction and Summarization*. Association for Computational Linguistics, 2008, pp. 17–24.

[26] Z. Han, X. Jiang, M. Li, M. Zhang, and D. Duan, "An integrated semantic-syntactic sblstm model for aspect specific opinion extraction," in *International Conference on Web Information Systems and Applications*. Springer, 2018, pp. 191–199.

[27] T. Joachims, "Svm-rank: Support vector machine for ranking," *Cornell University*, 2009.

[28] Y. Tay, M. C. Phan, L. A. Tuan, and S. C. Hui, "Skipflow: Incorporating neural coherence features for end-to-end automatic text scoring," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[29] D. Yogatama and N. A. Smith, "Bayesian optimization of text representations," *arXiv preprint arXiv:1503.00693*, 2015.

[30] Y. Qian *et al.*, "Improving native language (l1) identifation with better vad and tdnn trained separately on native and non-native english corpora," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 606–613.

[31] M. Dash and H. Liu, "Feature selection for classification," *Intelligent data analysis*, vol. 1, no. 3, pp. 131–156, 1997.

[32] T. Kohonen, "Correlation matrix memories," *IEEE transactions on computers*, vol. 100, no. 4, pp. 353–359, 1972.