http://doi.org/10.17576/apjitm-2022-1101-08
Asia-Pacific Journal of Information Technology and Multimedia
Jurnal Teknologi Maklumat dan Multimedia Asia-Pasifik
Vol. 11 No. 1, June 2022: 90 - 101
e-ISSN: 2289-2192

Received: 27 November 2021
Accepted: 10 January 2022
Published: 30 June 2022

ABLATION STUDY ON FEATURE GROUP IMPORTANCE FOR AUTOMATED ESSAY SCORING

JIH SOONG TAN
IAN K.T. TAN

ABSTRACT

Grading of written academic essays by humans requires significant effort. It is a time-consuming task and is vulnerable to human biases. Ever since the introduction of modern computing, this has been one of the many automations being explored. Researches in automated essay scoring have been on-going, where the majority of the researches in recent years are based on extracting multiple linguistic features and using them to build a classification model for automated essay scoring. The 3 main types of features used are lexical, grammatical, and semantic. In our work, we conducted an ablation study to discover the engineered features that has the weakest influence. We did this using a generic feature engineering and classification approach that was used by the winners of the Automated Student Assessment Prize (ASAP). This is to mitigate biases that may have addressed specific feature engineering or models. Our results show that a semantic feature called the prompt has been the weakest feature in influencing the models. From further investigations, this was due to it being over-fitted in the classification model.

Keywords**:** Automated Essay Scoring, Ablation Study, Feature Engineering, Semantic, ASAP.

INTRODUCTION

A focused piece of writing in responding to a prompt is defined as an essay. Essays are generally used in academic writing which determines the understanding of students based on their arguments. However, using human graders to evaluate an essay requires significant effort and time. More often than not, human grading is vulnerable to be biased and the outcome varies based on the events that happened in the human grader's life (Shermis & Burstein, 2003). Every human is different from each other which results in the diversity of styles to grade an essay and causes inconsistency in essay scoring. An automated essay scoring computing system ought to be capable of overcoming these human graders' foibles by being consistent and fair throughout the essay evaluation (Shermis & Burstein, 2003; Janda et al., 2019).

Automated Essay Scoring (AES) is an implementation that makes computers to be the graders which allows them to evaluate an essay written surrounding a prompt and provide a score to it. In the year 1966, Page shared the idea of an automated "grader" and invented an AES system called Project Essay Grade (PEG) (Page,1966). Since then, there have been a large number of innovations and new systems being developed in the AES field. Some of the most outstanding systems are, an improved version of PEG (Page, 1994), e-rater V2 (Attali & Burstein, 2006), and Intellimetric (Elliot, 2001). The majority of the existing systems are based on extracting multiple linguistic features that represent the quality of the essay and using them to produce a classification model for essay score prediction. Among all these systems, the linguistic features can be grouped into 3 main types of features: lexical, grammatical, and semantic features.

In this paper, we conducted an ablation study to evaluate the generic approach of feature engineering in AES. In the previous study by Shermis & Burstein (2003) and Ramest & Sanampudi (2021), they have reported that the key properties of a good essay are written around the given prompt, well-structured, smooth flow, good grammar application, suitable length, good spellings, and proper punctuation. Hence, we propose a feature influence study to find the weak point of current feature engineering on the generic approach of AES. The main benefit of our proposed experiment is to identify the weakness and for follow on research to address this area. We have implemented multiple learning algorithms for the classification models to discover the most influential and the least influential or the weakest component of the current feature engineering method.

RELATED WORK

Feature engineering allows the system to improve its understanding of the data by leveling up the abstraction of data continuously. For feature engineering in AES, there have been several recent works. Phandi et al. (2015) have implemented the Enhanced AI Scoring Engine (EASE) to extract features from essays and use the features to train an AES classification model based on Bayesian Linear Ridge Regression (BLRR) model. The authors managed to obtain a 0.7045 average Quadratic Weighted Kappa (QWK) score. The EASE engine provided 14 features with a total of 4 groups, namely length, part-of-speech (PoS), bag of words (BoW), and prompt. It is often being implemented by multiple works as the baseline feature engineering comparison to their own research experiment (Yang et al., 2020; Nguyen & Litman, 2018; Liu et al., 2019) as it is invented by one of the top 3 winners of the Automated Student Assessment Prize (ASAP) competition, so the engine has been witnessed to be robust.

Yannakoudakis et al. (2011) and Cummins et al. (2016) have applied the Robust Accurate Statistical Parsing (RASP) system proposed by Briscoe et al. (2006) in their work. The RASP system is a feature extraction technique that is similar to the EASE engine by grouping features into four main groups, namely lexical, part-of-speech, syntactic, and others. However, unlike EASE, the RASP system does not extract any prompt-specific or semantic features.

Coh-Metrix is a system invented by Graesser et al. (2004). It is a vast combination of software modules that extract features based on language, discourse, cohesion, and world knowledge (McNamara et al., 2010). It provides functions to extract 106 features with a total of nine main feature groups, namely semantic, discourse, syntactic, text descriptives, lexical diversity, text easability, connectives, word information, and referential cohesion. Latifi and Gierl (2020) implemented Coh-Metrix to extract features from the ASAP dataset and train a classification model based on a random forest algorithm. They obtained a 0.7 average QWK score, which is slightly worse in the QWK score than the EASE engine implemented by Phandi et al. (2015). Likewise, Chen and He (2013) have proposed extracting five main feature groups from the essay: lexical, syntactical, grammatical, contextual, and prompt-specific or semantic features similar to the EASE engine. In 2017, Eid and Wanas proposed feature engineering for AES classification models using lexical features only by gathering 22 lexical features from three other pieces of research. They reported their method improved the QWK by near 0.02 through the implementation of lexical features only.

Cozma et al. (2018) suggested a hybrid method of using string kernels and word embedding together to do feature engineering for AES prediction models. They hypothesized that the word embedding would help identify the features that the string kernel is short of. Also, string kernels can help identify similarities between words in a specific theme (Shawe-Taylor & Cristianini, 2004). Word embedding will help extract significant word vectors that represent the semantic or contextual meaning of the words (Mikolov et al., 2013).

Nguyen & Litman (2018) proposed a set of argumentative features for the argumentative AES task. They captured the 33 argumentative features based on existing works grouped in 5 main groups: argument component features, component label features, argument flow features, argumentative relation features, and argumentation structure typology features. The work managed to determine that the argumentative features can better improve argumentative AES classification as they manage to capture the semantic attributes within the essays.

In 2019, Janda et al. proposed an approach to extract 30 features with a total of three main groups, including syntactic, semantic, and sentiment. They have implemented several feature selection methods to filter out the weaker features, then input the selected features into a three-layer neural network to predict the output scores. They reported achieving a 0.793 average QWK score with near 0.1 improvements in QWK compared to BLRR, as Phandi et al. (2015) reported.

Later in 2019, Liu et al. (2019) proposed an AES system based on two-stage learning. In the first stage, the proposed model will calculate the semantic, coherence, and prompt-relevant scores based on deep neural networks. Then, in the second stage, the work will use the output of the first stage and feature engineered grammatical and lexical features as the inputs of a machine learning algorithm, XGBOOST, to perform the final AES classification task. The grammatical and lexical features engineered are similar to EASE: grammar error, essay length, word count, and vocabulary. Their work obtained an average 0.773 QWK score on the ASAP dataset.

To perform the AES task, Yang et al. (2020) implemented a fine-tuning pre-train language model, the BERT (Bidirectional Encoder Representations from Transformers) model. They believed that the model would consider deep semantics attributes from the essay, making their results superior to the EASE engine. However, the authors reported that the BERT model requires further effort to fine-tune the model to perform AES tasks for different essay prompts, which is ineffective for the main purpose of AES to reduce the effort to grade essays.

Generally, most AES's feature engineering primarily deals with three feature groups: lexical, grammatical, and semantic feature groups. The EASE engine implemented by Phandi et al. (2015) has robust results based on the related work review. Hence, it is a good investigation baseline as it extracts all three main feature groups: lexical, grammatical, and semantic refer to Table 1. We propose to base the evaluation on Phandi et al. (2015) to identify the influential strength of each feature group.

EVALUATION METHODOLOGY

Our investigation objectives are to focus on generic methods of feature engineering on AES apart from investigating the training part of the classification model in the AES system. Our ultimate goal of this experiment is to find the weaknesses of the handmade features and provide future direction to it.

DATA PREPROCESSING

For this experiment, we use essay set 2 from the ASAP competition-released dataset. We have only chosen one set of essays so that we can focus on our investigation. We have decided to apply set 2 because the selected set has the worst results in terms of quadratic weighted Kappa (QWK) scores reported in Phandi et al. 's (2015) paper. Essay set 2 has the prompt type of narrative essays in the format of story writing. Also, essay set 2 has 350 average word lengths and 1800 samples size. In terms of the scores, essay set 2 has the range of 0 to 6 score.

We implement the feature engineering on the essay set 2 by using the EASE engine. Over the years, EASE has been implemented by various researchers, and hence it is proven to be an appropriate platform for generic feature engineering method for AES system (Yang et al., 2020; Phandi et al, 2015; Latifi & Gierl, 2020;). EASE mainly grouped the feature into four groups which are length, part of speech (PoS), bag of words (BoW), and prompt. Features generated by EASE are shown in Table 1.

TABLE 1. Features generated by EASE

Feature Groups	Feature Names	Feature Descriptions	Main Feature Group
Length	chars	Count of characters.	Lexical
	words	Count of words	
	commas	Count of commas.	
	apostrophes	Count of apostrophes.	
	punctuations	Count of sentences ending punctuation symbols	
	avg_word_length	Average word length.	
Part-of-Speech (PoS)	POS	Count of bad PoS n-grams.	Grammatical
	POS/total_words	Ratio of bad PoS n-grams over total words count.	
Prompt	prompt_words	Prompt words count.	Semantic
	prompt_words/total_words	Ratio of prompt words count over total words count.	
	synonym_words	Synonym of prompt words count.	
	synonym_words/total_words	Ratio of synonym of prompt words count over total words count.	
Bag of Words (BoW)	unstemmed	Count of effective unstemmed unigram and bigram.	Lexical
	stemmed	Count of effective stemmed unigram and bigram.	

We based our evaluation methodology on the data preprocessing method reported by Phandi et al. (2015) to achieve the similar output they achieved with the EASE engine. We follow their steps to scale the scores to train the model and afterward will rescale the predicted score and rounded to the nearest score integer. For features preprocessing, we standardize length, PoS, and prompt features to a range of 0 to 1, and for BoW features are recalculated to log(1 + count). Train and test set split is required as the test set's scores are not provided in the dataset.

LEARNING ALGORITHMS

Apart from the learning algorithm of linear support vector machine (SVM) and Bayesian linear ridge regression (BLRR) that were implemented in Phandi et al.'s (2015) paper, we added Multinomial Naive Bayes (NB) in our evaluation methodology. Multinomial Naive Bayes, one of the typical Naive Bayes variants applied in text classification, is added as it is well known to deal with multinomial distributed data. As reported by Phandi et al. (2015), they picked BLRR as it has been proven to often provide good results in natural language processing jobs, and SVM regression as the comparison against BLRR. The implementation of the learning algorithm is coded in the Python (version 3.8) utilizing the scikit-learn library.

EVALUATION METRIC FOR LEARNING ALGORITHMS

To evaluate the trained models, we use QWK to calculate the agreement between two raters, the human rater and the trained models. We selected this evaluation metric as it is the official evaluation metric being implemented in the ASAP competition and proven to be a robust measurement for AES system by other work such as (Phandi et al., 2015; Yang et al., 2020; Latifi & Gierl, 2020; Janda et al., 2019; Cummins et al., 2016) that implements the same evaluation metric on the ASAP dataset.

EXPERIMENTAL SETUP

The original preprocessed data will be processed into eight sets and divided into two types of dataset: "only one feature group" and "excluding one feature group" to perform the ablation study. The "only one feature group" dataset is to determine the most influential feature among the features extracted from EASE. Vice versa, the "excluding one feature group" dataset is to determine the least influential feature or the weakest feature. "Only one feature group" dataset will consist of four sets of data that contain a single feature group of features, including "Only Length," "Only PoS," and "Only Prompt." Vice versa, "excluding one feature group" of the dataset will be consisting four sets of data that exclude a single feature group of features, including "Excluding Length", "Excluding PoS", "Excluding BoW" and "Excluding Prompt". The original preprocessed dataset will be the base comparison for the eight sets. Hence, there are a total of nine datasets.

We use 5-fold cross-validation on the processed datasets to split up the train and test set. We implemented 5-fold cross-validation on the ASAP dataset because the official test set is not released to the public. Also, 5-fold cross-validation ensures that every part in the dataset has the equal chance of appearing in the training and test set. The train to test set will be a 4 to 1 ratio, resulting in a 4-fold for the train set and 1-fold for the test set. The train sets will be taken to train the classification models using learning algorithms of Multinomial NB, SVM, and BLRR separately. Then, the trained AES classification models will be applied to predict the scores of the test sets. The QWK score will be calculated from the predicted and actual scores to measure the agreement between the human rater's scores and the predicted scores.

FEATURE INFLUENCE

We compare the differences in QWK score between the processed and the original datasets to identify the feature groups' influences on the prediction models. To measure the feature influence in "only one feature group" datasets, the higher the differences in QWK score between the original dataset and processed dataset indicate that the feature group has a lower feature influence in the model. The feature has a lower impact on the model and affects the

model's QWK score to deviate less from the original score. Vice versa, "excluding one feature group" datasets will have a stronger influence on the model for the excluded feature group if it has lower QWK score differences between the original and preprocessed datasets.

FEATURE SELECTION

The three feature selection techniques we applied to compare with previous results.

CHI-SQUARED

We are using the scikit-learn library to apply the SelectKBest class. We select the chi-squared (CHI) score function to measure the feature scores. CHI calculates the relationship of feature and target variables in a two-way contingency table. CHI is calculated as:

$$x^2(t,c) = \frac{N \times (AD - CB)^2}{(A + C) \times (B + D) \times (A + B) \times (C + D)} \quad (1)$$

Whereas in our case, each feature is the term t and score is the term c. Term A corresponds to the count of t and c occur simultaneously. Term B is calculated by taking the count of t that occurs when c not. Term C equal to the count of c occurs when t not, D is the count of both t and c does not occur. Term N is computed by summing up the total count of documents.

EXTRA TREE CLASSIFIER

We implement another feature selection technique from scikit-learn, an extra tree classifier to compute the feature importance value out of the features. The extra tree classifier put up multiple trees and randomly split nodes using random subsets of features which brings up its benefit of randomizing characteristics for numerical inputs (Sharaff & Gupta, 2019). This may be effective for the experiment because the generated features are all numerical inputs.

CORRELATION MATRIX WITH HEATMAP VISUALIZATION

We use the correlation matrix to compute the scores of the relationship that exists between features and target variable. It is a classic feature selection method for machine learning methods such as logistic and linear regression due to its' ability to eliminate variables that has the weak relationship to the target variable (Bisong, 2019).

RESULTS AND DISCUSSION

QWK SCORES RESULT FOR COMPARISON

We determine the QWK scores of the original "all features" dataset, the four "only one feature group" datasets, and four "excluded one feature group" datasets based on the trained model of NB, SVM regression, and BLRR. The trained model results based on the original dataset are shown in Table 2, where it shows the BLRR model outperforms the rest of the models as suggested in Phandi et al. (2015) paper.

TABLE 2. All Features Experimental Result

Feature Used	QWK Score		
	NB	SVM	BLRR
All features	0.517	0.601	**0.626**

The table 3 contains the trained classification models' "only one feature group" results, where the strongest feature group influence is bold-faced, and the weakest feature group influence is underlined. We see that the length feature group is the strongest feature group in NB and BLRR models, as it has the least differences in QWK score with the "all features" trained model.

TABLE 3. Only One Feature Group Experimental Result

Feature Used	QWK Score		
	NB	SVM	BLRR
Only Length	**0.504**	0.580	**0.604**
Only PoS	0.483	0.584	0.536
Only BoW	0.000	**0.590**	0.574
Only Prompt	0.246	0.569	0.543

The table 4 contains the trained classification models' results for "excluded one feature group" are tabulated in table 4. Similarly, the strongest feature group influences are bold-faced, and the weakest feature group influences are underlined. From table 4, we can again see the length feature group to be the strongest feature group among all. However, the prompt feature seems to be lacking here. The QWK score of "excluding prompt" in SVM and BLRR compared to "all features" shows it's overfitting the trained model. By overfitting, it means the prompt feature has worsened the models.

TABLE 4. Excluded One Feature Group Experimental Result

Feature Used	QWK Score		
	NB	SVM	BLRR
Excluding Length	**0.444**	**0.565**	**0.601**
Excluding PoS	0.511	0.583	0.617
Excluding BoW	0.546	0.599	0.604
Excluding Prompt	0.494	0.636	0.657

To identify the reason for overfitting on the prompt feature, we have looked into the EASE engine. The EASE engine generates prompt features in a very simple way. The engine tokenizes the prompt or essay topic into prompt words by implementing Python Natural Language ToolKit (NLTK). Then, it uses the WordNet corpus in NLTK to identify the synonym of the prompt words being written in the prompt. The EASE engine counts the number of the synonym of prompt words and prompt words being written in each essays and calculate its ratio.

We hypothesize the main reason for the least influential and overfitting of prompt feature groups in the models is its weakness in identifying semantic attributes from the essays and prompt. Semantic attributes are attributes corresponding to the contextual meaning of a word or a group of words (Janda et al., 2019). It is essential for essay evaluation to be written

around the prompt or the essay topic semantically (Norton, 1990). Therefore, we hypothesize that the EASE engine considers all part-of-speech types in the essays and the prompt, which leads to noisy data in prompt features and overfitting the model. Part-of-speech types such as conjunction, adposition do not contain any contextual meaning, which may add noise to the dataset.

In addition, we hypothesized that EASE's method of extracting semantic attributes from essays and the prompt is too brief and can be further improved in the future. It only considers individual words rather than phrases or sentences, which make it impossible to detect if a sentence or essay is starting to digress. The essays may contain prompt words or synonyms for prompt words, but the topic might not be connected between phrases or sentences. Miltsakaki & Kukich (2000) have reported that the disconnection of the topic in between parts of the essay evidence that the disconnected part is disjointed from the other parts of the essay, and this would result in topic digression. Hence, it is crucial to keep the coherence between phrases or sentences to make the written content semantically meaningful.

FEATURE SELECTIONS RESULT FOR COMPARISON

CHI-SQUARED

The results of the CHI feature selection technique are tabulated in Table 5, where the features are ranked descending based on feature scores (higher the feature score, closer the relationship to the target variable). As expected, the length feature is the most important feature as it has two features (chars, words) out of five that took the top two highest feature scores. Likewise, the PoS feature (POS, POS/total_words) ranked somewhere in the middle among the features. In addition, BoW features (unstemmed, stemmed) outperform PoS and prompt features in terms of the feature scores. As we anticipated based on the results in feature influence, the prompt feature has two features (synonym_words/total_words, prompt_words/total_words) out of four ranked at the bottom of the table. This proves the experiment we have done in section feature influence is correct.

TABLE 5. Chi-squared Experimental Results

Features	Feature Scores
chars	307333.663565
words	56566.133705
unstemmed	48743.458825
stemmed	47796.792412
prompt_words	26154.560390
synonym_words	11129.269480
commas	4166.392504
punctuations	2480.382235
POS	1152.672316
apostrophes	903.672316
avg_word_length	2.114962
POS/total_words	1.020496
synonym_words/total_words	0.778495
prompt_words/total_words	0.126973

EXTRA TREE CLASSIFIER FEATURE SELECTION

We generated the result of extra tree classifier feature selection and plot it in a bar graph as shown in Figure 1, where the features are ranked ascending based on the feature importance score. Similar to CHI and the result in the feature influence experiment, the extra tree classifier ranked the prompt feature as the least important feature among all as the two (synonym_words/total_words, prompt_words/total_words) out of four features are ranked at the bottom of the graph. However, the most important feature in the extra tree classifier is the BoW features (stemmed, unstemmed).

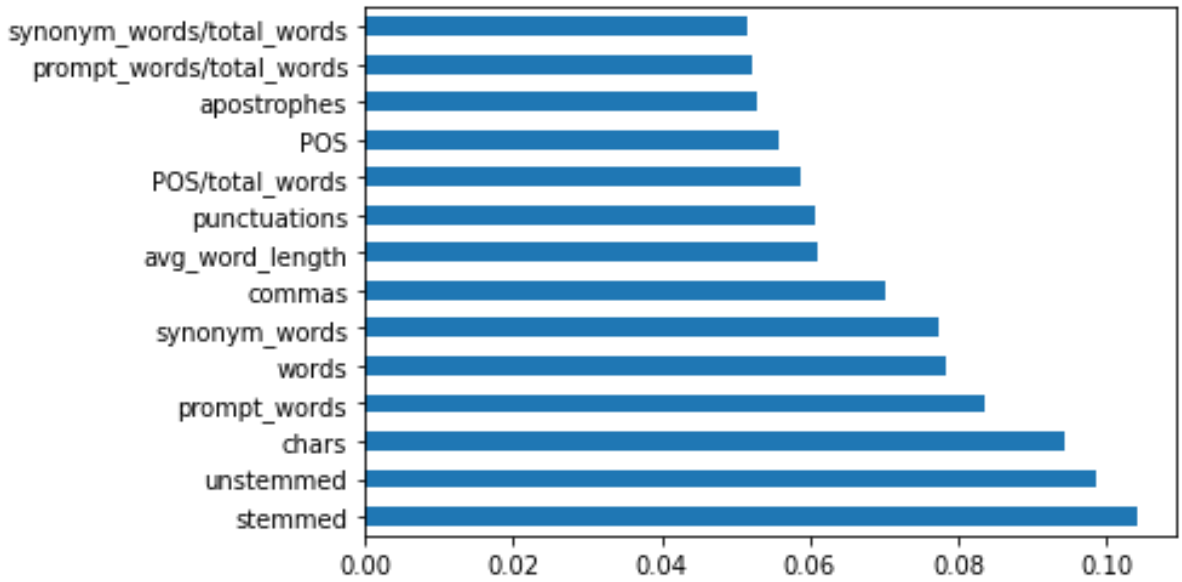

FIGURE 1. Extra Tree Classifier Experimental Result

CORRELATION MATRIX

The result of the correlation matrix is generated and visualized in a heat map as shown in Figure 2, where the higher or greener the value is, the higher the correlation. The most correlated variable is similar to the extra tree classifier, it shows the BoW features (stemmed, unstemmed) to be most correlated with the target variable, score by average correlation score at 0.7. Unlike the extra tree classifier, the least important feature in the correlation matrix is the PoS feature where the average correlation score of PoS features is the lowest at 0.03. However, the prompt feature has the second-lowest average correlation score at 0.25575 which is still acceptable to match our result in the feature influence experiment.

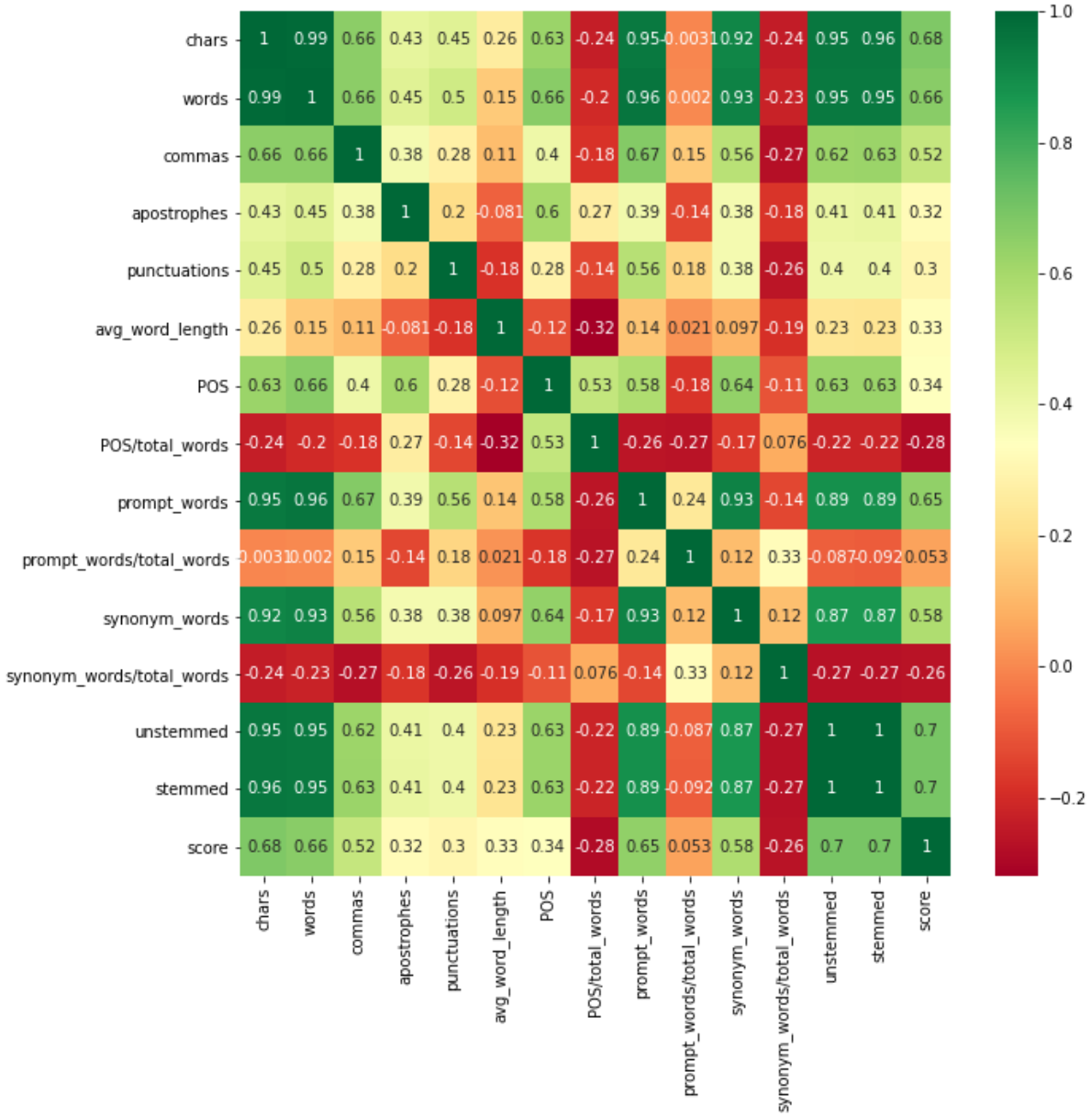

FIGURE 2. Correlation Matrix with Heatmap Visualization.

CONCLUSION & FUTURE WORK

We have experiments to study the weaknesses of the generic method of feature engineering in AES using the ASAP dataset through the EASE engine. We propose to evaluate the four main feature groups based on the EASE engine by using "only one feature group" sets and "excluding one feature group" sets, then compare their QWK score with the "all features" set. As the comparison between the sets, our work has shown that the prompt feature is the weakest feature among the four main features groups. The "excluding one feature group" set has represented that the QWK scores of SVM and BLRR without prompt features are better in performance. Hence, the experiments show that the prompt feature group is overfitting in the dataset. To make sure what we did to rank to feature influence is correct, we have done multiple feature selection techniques to compare with the ranking we done. Thus, it provides accurate information and enough details for us to work on the new prompt feature engineering. As such, we can work on researching the new prompt feature in the future.

REFERENCES

Attali, Y. and Burstein, J., 2006. Automated essay scoring with e-rater® V. 2. *The Journal of Technology, Learning and Assessment*, *4*(3).

Briscoe, T., Carroll, J.A. and Watson, R., 2006, July. The second release of the RASP system. In *Proceedings of the COLING/ACL 2006 interactive presentation sessions* (pp. 77-80).

Bisong, E., 2019. *Building machine learning and deep learning models on Google Cloud Platform* (pp. 7-10). Berkeley: Apress.

Chen, H. and He, B., 2013, October. Automated essay scoring by maximizing human-machine agreement. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (pp. 1741-1752).

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

Cozma, M., Butnaru, A.M. and Ionescu, R.T., 2018. Automated essay scoring with string kernels and word embeddings. *arXiv preprint arXiv:1804.07954*.

Elliot, S., 2003. IntelliMetric: From here to validity. Automated essay scoring: A cross-disciplinary perspective, pp.71-86.

Eid, S.M. and Wanas, N.M., 2017, November. Automated essay scoring linguistic feature: Comparative study. In 2017 Intl Conf on Advanced Control Circuits Systems (ACCS) Systems & 2017 Intl Conf on New Paradigms in Electronics & Information Technology (PEIT) (pp. 212-217). IEEE.

Graesser, A.C., McNamara, D.S., Louwerse, M.M. and Cai, Z., 2004. Coh-Metrix: Analysis of text on cohesion and language. Behavior research methods, instruments, & computers, 36(2), pp.193-202.

Janda, H.K., Pawar, A., Du, S. and Mago, V., 2019. Syntactic, semantic and sentiment analysis: The joint effect on automated essay evaluation. IEEE Access, 7, pp.108486-108503.

Liu, J., Xu, Y. and Zhu, Y., 2019. Automated essay scoring based on two-stage learning. arXiv preprint arXiv:1901.07744.

Latifi, S. and Gierl, M., 2021. Automated scoring of junior and senior high essays using Coh-Metrix features: Implications for large-scale language testing. Language Testing, 38(1), pp.62-85.

Miltsakaki, E. and Kukich, K., 2000. Automated evaluation of coherence in student essays. In Proceedings of LREC 2000 (pp. 1-8).

McNamara, D.S., Louwerse, M.M., McCarthy, P.M. and Graesser, A.C., 2010. Coh-Metrix: Capturing linguistic features of cohesion. Discourse Processes, 47(4), pp.292-330.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. and Dean, J., 2013. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).

Norton, L.S., 1990. Essay-writing: what really counts?. Higher Education, 20(4), pp.411-442.

Nguyen, H. and Litman, D., 2018, April. Argument mining for improving the automated scoring of persuasive essays. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 32, No. 1).

Page, E.B., 1966. The imminence of... grading essays by computer. The Phi Delta Kappan, 47(5), pp.238-243.

Page, E.B., 1994. Computer grading of student prose, using modern concepts and software. The Journal of experimental education, 62(2), pp.127-142.

Phandi, P., Chai, K.M.A. and Ng, H.T., 2015, September. Flexible domain adaptation for automated essay scoring using correlated linear regression. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (pp. 431-439).

Ramesh, D. and Sanampudi, S.K., 2021. An automated essay scoring systems: a systematic literature review. Artificial Intelligence Review, pp.1-33.

Shermis, M.D. and Burstein, J.C. eds., 2003. Automated essay scoring: A cross-disciplinary perspective. Routledge.

Shawe-Taylor, J. and Cristianini, N., 2004. Kernel methods for pattern analysis. Cambridge university press.

Sharaff, A. and Gupta, H., 2019. Extra-tree classifier with metaheuristics approach for email classification. In Advances in Computer Communication and Computational Sciences (pp. 189-197). Springer, Singapore.

Yang, Y. and Pedersen, J.O., 1997, July. A comparative study on feature selection in text categorization. In Icml (Vol. 97, No. 412-420, p. 35).

Yannakoudakis, H., Briscoe, T. and Medlock, B., 2011, June. 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 (pp. 180-189).

Yang, R., Cao, J., Wen, Z., Wu, Y. and He, X., 2020, November. Enhancing Automated Essay Scoring Performance via Cohesion Measurement and Combination of Regression and Ranking. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings* (pp. 1560-1569).

Jih Soong Tan
Priority Dynamics Sdn Bhd
jsoong@prioritydynamics.com

Ian K.T. Tan
Heriot-Watt University Malaysia
i.tan@hw.ac.uk