2022 IEEE 7th International conference for Convergence in Technology (I2CT)
Pune, India. Apr 07-09, 2022

A Hybrid Approach Towards Automated Essay Evaluation based on Bert and Feature Engineering

Shreya Prabhu
Computer Science
PES University
Banglore, India
shreya.pr611@gmail.com

Kara Akhila
Computer Science
PES University
Banglore, India
karaakhila@gmail.com

Sanriya S
Computer Science
PES University
Banglore, India
sanriyasan@gmail.com

2022 IEEE 7th International conference for Convergence in Technology (I2CT) | 978-1-6654-2168-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/I2CT54291.2022.9824999

***Abstract*—Educational institutions often assess a student's critical thinking and communication skills based on essay responses. Manual Evaluation is time-consuming, and there may be wide variations when multiple evaluators rate batches of essays. In the last few years, the automated grading of essay scripts has emerged as a new area of research. Most studies essentially focus on visible attributes such as length, vocabulary, sentiment or spelling. The use of neural networks requires the conversion of text into some vector representations. However solely using handcrafted attributes or text encodings implies primarily operating on word granularity. On the other hand, Transformers can handle dependencies between words of the text. In this paper, we propose a hybrid model that can capture the interaction of words in the essay using the BERT self-attention transformer, along with handcrafted syntactical features. While previous studies have built individual models for every essay topic, our model has been incrementally trained on multiple essay topics to test its generalizability. The validation of the model uses quadratic weighted kappa to compare human-rated scores and model scores.**

Keywords - essay, automated grading, BERT, self-attention, kappa

I. INTRODUCTION

An Automated Essay Scoring (AES) system aims to provide fast, effective and affordable solutions for automated grading of essays. This system is necessary given that essay writing is not only deployed as a part of school assessments but is also a crucial step in gaining admission to higher institutions of study. Manual evaluation becomes tedious as the number of participants increases. There can be variations in the manner in which two raters assign scores to the same essay. While some raters may be impressed by the use of extensive vocabulary, others may focus more on the content of the text, even if the language used is very simple. Evaluators may get fatigued reading scripts and could end up relaxing their evaluation. Thus, an automated solution that evaluates essays uniformly and consistently is required. It would save time for both the institution faculty as well as the students.

One of the main challenges faced for an AES system is the vast domain of topics that exist. The prompts provided to the students can fall under several categories. For the dataset used in this study, the given prompts are either source-based or descriptive. The uniqueness of each student's response is another matter to be considered. There is no fixed formula or model answer to represent the best-written essay. We aim to address these constraints by combining multiple prompts of our dataset. This approach would help our model cover as many aspects of writing as possible under several topics.

There has been a large amount of ongoing research in this field. One technique of analyzing text scripts focuses on assigning numerical values to attributes such as length, vocabulary, sentiment or spelling and understanding how much these weigh in to give a score. However, most of these fail as they only look at extracting granular features. It is necessary to understand the interaction of the subjects, objects and verbs of all the sentences to capture the true meaning of the essay. Recent developments in Natural Language Processing (NLP) have focused on the benefits of using pre-trained models on related tasks. One such example is the Bidirectional Encoder Representations from Transformers (BERT) model, which uses self-attention to relate words in the text and compute a representation of it. It provides weights to the text embeddings to include contextual relevance.This model takes care of the aforementioned statement of understanding the relationship between one section of the essay and another. The dataset used in this study consisted of school level essays.Grammatical construction and misspellings are frequent aspects of the rubrics for scoring. Hence, syntactical errors were considered important factors in assigning scores. We propose a hybrid model that can capture the interaction of words in the essay using the BERT self-attention transformer, along with handcrafted syntactical features.

II. RELATED WORK

The evaluation and scoring of essays in many prior automated essay scoring systems were based on carefully crafted features. In one such system [1], authors considered attributes that represent the style of writing. Some of these are imagery, emotive effectiveness, age of acquisition and the beauty of words. Many of these features were dependent on the age of the writers and prompt. This model focused only on the vocabulary and emotion of the essay and did not yield good results. Another such model used novel graph-based semantic features with syntactic and sentimental features [2]. Graph-based representations help discover associations and patterns within a chunk of data. For coherence, the semantic similarity was computed between all pairs of sentences to analyze the overall coherence. Traits such as eccentricity, closeness centrality, minimum and maximum spanning trees describe relations between sentences. The minimum and maximum spanning trees derived the strongest and weakest similarity connections in the essay. Graphs with higher average centrality values had more cohesion associated with the essay. Syntactical attributes included by the authors were unique parts of speech used, misspelt words, average sentence length, subject-verb agreement and more.

Proposing fixed baseline features that would determine the score given to an essay requires a lot of trial and error, as each rater may grade an essay based on different benchmarks that they consider as important. Some may pass over syntactic errors as long as the content of the easy strikes well. [3] proposed that instead of relying on feature engineering, a system that automatically learns the representations between the words in the essay is needed in order to compute the quadratic kappa score that is needed for the assignment. Contrary to models that used only manually crafted features, authors of [3] used neural networks to effectively encode the information required for essay evaluation and learn the complex patterns in the data through non-linear neural layers. The convolutional layer provided n-gram level information to potentially capture local contextual dependencies in the essay and the Long Short Term Memory (LSTM) layer extracted long term dependencies. Finally, a linear layer with sigmoid activation mapped its input vector generated by the mean-over-time layer to a scalar value to calculate the score. The authors claimed that this method effectively utilized essay content to extract the required information for scoring essays. However, the network learned to take essay length into account and assigned low scores to those with fewer than 50 words, regardless of the content. In terms of quadratic weighted Kappa, the best method (ensemble model) outperformed the baseline by 5.6 percent.

In paper [4] the authors proposed the structure of the Transformer, a novel architecture that solves sequence-to-sequence tasks while handling long-range dependencies with ease. While the encoder portion takes in a tokenized input and maps it to an abstract representation, the decoder takes that continuous representation step by step to generate an output while being fed previous results. They further define self-attention as a mechanism that relates different positions of sequences to compute a representation of it. Each encoder layer has a self-attention mechanism and a fully connected feed-forward network. The authors further refined the self-attention layer by adding "multi-headed" attention. It expands the ability of the model to focus on different positions and gives the attention layer multiple "representation subspaces". The decoder had an additional sub-layer that performed multi-head attention over the output of the encoder stack.

The authors of the paper [5] proposed the architecture of BERT (Bidirectional Encoder Representations from Transformers). BERT's model architecture is a multi-layer bidirectional Transformer encoder based on the original implementation described in [5]. The model comprises 12 layers, 12 attention heads, a hidden size of 768 and 110 million parameters. The BooksCorpus and English Wikipedia corpora were used for pre-training. As a part of pre-processing, the first token of every sequence was marked as a special classification token called CLS. The states of the entire sequence, on training, finally aggregate into this token. Furthermore, the sentences were distinguished with the help of a special token called SEP. The CLS layer is fed into an output layer for the specific task. In this study, it would be classification. The tasks used for pre-training were masked language modelling and next sentence prediction. BERT distinguishes itself from other language models as it uses bidirectional representations. It jointly conditions on the left and the right contexts in all layers. Furthermore, it has a unified architecture across different tasks. For fine-tuning, the BERT model can be initialized with the pre-trained parameters, and all of the parameters are updated using labelled data from the downstream tasks. The paper describes instances of fine-tuning BERT that include sentence pair paraphrasing, hypothesis-premise pairs in entailment, question-passage pairs in question answering, and a degenerate text-null pair in the case of text classification or sequence tagging. Each downstream task has separate fine-tuned models, despite being initialized with the same pre-trained parameters. Thus, BERT is now being used in most problems in NLP as it is trained on a rich set of corpora, and easily adapts to the given problem set.

In [6], the authors use BERT for essay evaluation. The model takes a batch of essays as input. [CLS] is added at the beginning of each sample so that the encoded representations are the output vectors mapping to it. BERT is used to generate hidden representations of the text that are passed through a fully connected neural network to learn the features of the essay. The scores generated by the network are constrained by regression and ranking loss which are optimized jointly with the dynamic combination. The limitation faced in this study was that BERT limits the length of each input text, which in turn restricts the size of essays. Absence of the remaining portion of the essay while evaluating could lead to faulty results.

Authors of [7] describe a model that can integrate both handcrafted features as well as use the encoded features from transforming an essay into the vector form. Semantic scores are used to understand the underlying deep semantic meanings of the essay. The authors make use of LSTMs to convert the essays to vector embeddings as an input to the LSTM and map the final score to the range of [0,1]. The Coherence score marks the relation between different paragraphs that comprise the total essay. Coherence scores for well-organised essays are high, and the LSTM model is used again to calculate the score. Prompt Score analyses the connection between the essay contents and the prompt provided. Some other features taken into account by the authors include essay length, spelling and grammatical errors. Ensemble Learning methods are used to combine the results of different models to strengthen the final result. Decision trees are also used together to build an ensemble model. One such example is XGBoost, which makes use of a gradient boosting framework. Boosting is a framework where each decision tree attempts to build upon the errors made by the previous to improve final performance. The final score for this entire model is taken as Handcrafted Features Score, Semantic, Coherence and Prompt Score as parameters to the XGBoost framework.

While feature assessment is an implicit part of evaluating the quality of an essay, it cannot be easily generalized over different essay types (narrative/ comprehension/ argumentative). Transfer Learning via pre-trained models can help in improving the quality of vector embeddings generated in neural network models. This can enhance semantic and sentiment analysis of the essays. BERT is a significantly effective model that is used to improve the performance of downstream natural language processing tasks. Self-attention is useful to capture the relations of words within all essays, conjunction words and key concepts in essays. Modeling dependencies between a given words and the remaining words of the text is possible with self-attention. Our work proposes a hybrid model that combines

scores generated by BERT with additional handcrafted features related to Parts of Speech (POS) tags along with regular syntactic errors.

III. Dataset

The Automated Student Assessment Prize (ASAP) Dataset[1] has been used to develop the AES model. The dataset consists of over twelve thousand essays across eight different prompts. Each prompt set has between one and two thousand essay responses, each approximately 150 to 550 words in length. Some of the essays are dependent upon source information. Essays with prompt IDs 3, 4, 5 and 6 were source article based essays while 1, 2, 7 and 8 were descriptive essays. The responses are by students in grade levels ranging from Grade 7 to Grade 10. The attributes of this dataset are prompt number, a unique essay ID, two human rater scores and a resolved score across two domains. All essays have a resolved score in domain one, which we used to test the performance of the scoring engine. In our implementation, each prompt set's scores were scaled to [0,10] for uniformity. Around eighty percent was taken to train the Bert model.

A. External data samples

The model was also tested on external datasets. Scholastic Assessment Test (SAT)[2] and Graduate Record Examinations (GRE)[3] sample essays were used as unseen data. The SAT[1] dataset had two prompts, each having 8 essays. The score range of the 16 essays varied from three to twelve. The GRE[2] dataset consisted of two different prompt samples, each having 6 essays. The score range of the 12 essay samples varied from 1 to 6.

IV. Proposed Methodology

A. Training BERT

BERT is one of the most successful pre-trained transformer models available for NLP tasks. BERTForSequence from the Simple Transformers library is one such model used for classification and regression tasks. It provides a sliding window that overcomes the limitation of 512 words that BERT can process. This library is computationally intensive and must run on a Graphics Processing Unit (GPU). The trained model is then used to evaluate the remaining portion of the dataset.

B. Feature Generation

The second module involved estimating the diversity in vocabulary, wordy sentences, count of the most frequently used word, and syntactical errors. The second portion of the dataset would form the hybrid dataset consisting of the generated features and the bert score. A python wrapper for LanguageTool, an open-source grammar tool, was used to detect spelling errors, grammatical mistakes and punctuation errors. The Natural Language Toolkit (NLTK) was to generate Parts of Speech (POS) tags. The NLTK word tokenizer converts individual words of the essays to tokens. Stop words are common words such as articles, prepositions, pronouns, conjunctions that do not add much information to the text, which are filtered out. The tokens were then tagged. Coordinating conjunctions join two sentences. These were counted along with the number of unique parts of speech. [3] shows that long sentences and the frequent use of the same parts of speech indicate poor writing skills. The Type to Token Ratio (TTR) is a ratio of the total number of unique words present in the essay (types) to the total number of words (tokens). TTR captures vocabulary variation and is a measure of lexical diversity. A higher TTR would lead to a better score. The count of the most frequent word was also an attribute. A higher count represents a lack of extensive vocabulary and would therefore lead to a lower score.

C. Linear Regression

The outputs from the above two modules i.e, the BertForSequence score and the generated features are merged together to form a hybrid dataset. This dataset was split into train and test subsets. The supervised learning algorithm, Linear Regression was used to model the relationship between all the generated variables to the resolved score. Finally, the output values were compared to the original targets with the kappa and Root Mean Squared Error (RMSE) as the performance metrics.

D. Incremental Data Addition

The final step in our study was to check whether the generated features and BERT self-attention scores were powerful enough to be applied on both source-based and descriptive essays when combined. When the model is trained on essays under the same topic, it might learn to map scores according to common keywords. In the ASAP Dataset, different topics have completely varying contexts and expect diverse styles of responses. Furthermore, these also belong to students of different grade levels. When the responses from different sets are combined and shuffled, the model will judge the quality of writing. Thus, after testing on prompt set 1, set 2 was added to the dataset and tested again. This was followed for the remaining sets.

E. Performance Metric

The evaluation metric used to determine the performance of the proposed approach is quadratic weighted kappa. Kappa is a measure of the agreement between two raters or inter-rater agreement. It does not measure the absolute difference between the two raters and that makes it an ideal evaluation metric. Here, it is used between the predicted scores and the resolved scores from the dataset. The predicted score is rounded off to a point one decimal precision.

V. Experimental Results

The model was tested taking one or more prompt sets at a time. The predicted scores were tested against the resolved score in the dataset. Kappa and Root Mean Square Error (RMSE) were used as evaluation metrics, as shown in Table I. The model was then validated with external datasets, i.e SAT and GRE sample essays. These essays are used as unseen data and the results obtained are recorded in Table II.

TABLE I. Results on ASAP Dataset

Prompts	Performance Metrics	
	Kappa	***RMSE***
1	0.731	0.879
1-2	0.747	1.071

[1] https://www.kaggle.com/c/asap-aes
[2] https://collegereadiness.collegeboard.org/sample-questions
[3]https://www.ets.org/s/gre/accessible/gre_practice_test_3_writing_responses_18_point.pdf

1-3	0.771	1.25
1-4	0.797	1.362
1-5	0.769	1.44
1-6	0.811	1.368
1-7	0.79	1.38
1-8	0.802	1.321
Source Based (3,4,5,6)	0.808	0.994
Descriptive (1,2,7,8)	0.786	1.458

TABLE II. TESTING RESULTS ON EXTERNAL DATA SAMPLES

Metric	**Datasets**		
	ASAP Dataset	***SAT***	***GRE***
Kappa	0.802	0.554	0.397
RMSE	1.321	2.597	3.537

VI. DISCUSSION

Developing an AES to grade essays is vital to provide uniformity during assessment and save time spent in manual evaluation. Essay topics are of multiple types and subjects. There are different expectations from a source article based essay and a descriptive essay. Transformers have proven to be very successful in NLP tasks. BERT is one such self-attention based transformer. In transfer learning, pre-trained models can be used with only the last layer of the network retrained according to the problem statement. Feature engineering also has its merits, provided the right kinds of features have been derived. Transitioning towards an all-around grading model requires continual learning. More data would help optimize the model and improve its performance as time progresses.

VII. CONCLUSION

We have described an approach to building an AES system that incorporated both transformers as well as the use of feature generation. As the dataset consisted of school level essays, elementary features such as grammatical and spelling errors were used along with parts of speech characteristics. To cover as many aspects of writing over multiple topics and essay types, we combined the BERT score yielding relationships between sequences of the essay and the basic syntactic characteristics. There are expected variations as we add more prompts during training. But the results shown display consistency in values of Kappa and RMSE metrics. While the Kappa value remains in the range of 0.7 and 0.8, RMSE is close to 1. This proves that choosing the handcrafted attributes common to all well-written scripts, and the use of BERT gives a satisfactory performance. We can conclude that incremental training on similar topics and types might prove better than totally random topics.

ACKNOWLEDGMENT

We would like to express our gratitude to Prof. Suresh Jamadagni, Department of Computer Science and Engineering, PES University, for his continuous guidance, assistance, and encouragement throughout the development of this project.

REFERENCES

[1] S. Song en J. Zhao, “Automated essay scoring using machine learning”, Stanford University, 2013.

[2] H. K. Janda, A. Pawar, S. Du, en V. Mago “Syntactic, Semantic and Sentiment Analysis: The Joint Effect on Automated Essay Evaluation” , IEEE Access, vol 7, bll 108486–108503, August 20, 2019.

[3] K. Taghipour en H. T. Ng, “A neural approach to automated essay scoring”, in Proceedings of the 2016 conference on empirical methods in natural language processing, 2016

[4] A. Vaswani et al., “Attention is all you need”, in Advances in neural information processing systems, 2017, bll 5998–6008, December 6, 2017

[5] J. Devlin, M.-W. Chang, K. Lee, en K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding”, arXiv preprint arXiv:1810. 04805, May 24, 2019

[6] R. Yang, J. Cao, Z. Wen, Y. Wu, en X. He, “Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking”, Findings of the Association for Computational Linguistics: EMNLP, vol 2020, bll 1560–1569, 2020.

[7] Liu, Jiawei, Yang Xu, and Yaguang Zhu. "Automated essay scoring based on two-stage learning." arXiv preprint arXiv:1901.07744 (2019).