Grade It: A Quantitative Essay Grading System

DR. SINDHU
Department of Computing
SRM Institute Science and Technology
Kattankulathur, Tamil Nadu

ROOPCHAND REDDY VANGA
Department of Computer Science and Engineering
SRM Institute Science and Technology
Kattankulathur, Tamil Nadu
rv8364@srmist.edu.in

M.S. BHARATH
Department of Computer science and Engineering
SRM Institute Science and Technology
Kattankulathur, Tamil Nadu
Bs6396@srmist.edu.in

Abstract*—Automated writing evaluation system that employs automated essay scoring technologies Generates a rating to the writings which help students with self-assessment. It provides an efficient and easy way for grading Essays, which usually takes abundance of time to be evaluated by human graders, this can be greatly useful for educational institutions like schools and colleges. It provides ratings for essays based on the grammatical errors and topic relevancy by using specific tools implemented in the auto grading system. The generated result will help students grading essays which helps students in self-assessment, this will be useful for the student to understand the mistakes that he makes which will be pinpointed by the system in matter of time, not only that the system also displays the strengths of the user, system will improve the accuracy compared to already existing automated grading system. The system uses scores from multiple aspects (handcraft features, coherence score, Prompt-relevant score and semantic score). these key features illustrated above will provide enough information on the essay for the student

1.INTRODUCTION

With the day by day increase in the data availability, the time to extract information needs to evolve more and more efficiently. Likewise, the time to grade assignments takes a lot of time to do manually, especially in educational institutions where the teachers / professors have to grade for entire classes. This takes a lot of time not only for the graders but also

students to receive the graded scores and the remarks in with a lot of exams and assignments spawning everywhere in online during the covid period, it was physically impossible for both teachers and students to be able to conduct and attend physical mode of classes, these situations brought us closer to usage of online platform for everything, from classes to conducting exams. One specific topic that must be discussed is the grading system used, without the presence of pen-paper mode of examination most educational institutions had to start using AI grading for students, In the case of essays the grading is even harder as it takes a lot of time for the professor to grade it as he needs to check for grammatical mistakes, spelling mistakes, concurrency etc. The time taken for grading these essays could be utilized for other productive works.

Currently, AWEs are mostly used by researchers and scholars because of their technicality and low accuracy. Most open-source AWEs used in the industry have an average accuracy of 60%. feedback provided by most existing system only includes information such as mistakes in the writing. AWEs are widely used in competitive exam grading such as ETS' TOEFL, GRE etc. Our model primarily aims on improving accuracy if existing model, making it more accessible for educational institutions on a regular basis.

Our project mainly focusses on using the grading system to process more greatly and efficiently results, giving the students the necessary information regarding the mistakes and their strengths using the grading system that we developed, the system uses natural language processing using multiple tools specifically for analysing the sentences and grade it individually. Some of the key tools being used are structure checking, spelling checking, punctuation mark, stop-word removal etc. The system is trained by collecting pre-existing data and grading the models based on it. It also checks for relevance between the written essay and topic for grading. The model's accuracy will be gauged on the basis of its similarity with real-world manual evaluation by professor and instructors.

It starts with abstract which provides a gist of the topic and later accompanied by the introduction which gives more detailed explanation on the topic and reasons to choose, we get an overview of the steps and component used in the creation the system. We discuss about current existing system and their short comings which we will try to resolve in the fourth coming, where we discuss the methodology and experimentation along with results.

REFERENCE ID	FEATURES	ALGORITHM USED	UNADDRESSED FEATURES
[12]	Use of handcrafted features like Checking punctuation, Missing letters, spelling mistakes etc.	Stemming and String based Algorithm	In terms of Coherence and semantics it's not the best model.
[15]	Uses text vectorization and Semantic analysis	Bi-LSTM-CRF model.	Doesn't detect adverbial essay and essays with permuted sentences.
[8]	Scores different aspects of an assay and merges them	CNN+LSTM	Data used in this model is scarce, hence hard to find.
[18]	Focusses on critical words and analyse the logic semantic relationship	LSTM Model	Variety of dataset available, hence hard to go with particular
[10]	uses the BERT for embedding, along with handcrafted features to predict the score.	Supervised Learning Algorithm	BERT model is slow to train due to weight to update, also its an expensive model

2. RELATED WORK

The model is that is being developed is with the information gathered from other systems unaddressed features, advantages and their outcomes. An essay system developed in Saudi Arabian researchers [12] provided a wide variety of handcrafted features such as punctuation, checking for missing letters, spell check etc, they were successfully able to grade the essay with decent results but the model lacks coherence and semantics grading of the essay. Hence, handcrafted features were referred from their research work.

Another system had a good coherence grading, but the data set provided by the system was very less and the to run a essay system like the one that is provided by the institutions the system need a dataset that is huge. Most of the systems or projects used the vectorization as the main feature in the grading system, as it processes human language and understands meaning and context, vectorizations has been in use since the first computers were built it has been successful in various domains.

A hybrid model [8] uses BERT* model, it is mainly used in semantic analysis. which has the ability to process large amount of text and language. Which is very helpful while learning of pre -trained models, but this model also has its disadvantages as the model require large amount of data set to train and weight to update. These features will provide a better grading system for the user but then they also have their disadvantages, the purpose of this review is to identify the drawbacks of other working models and try to update the current model with the features and try to provide a more efficient way of grading.

3. GRADE IT: METHODOLOGY

In this section, the proposed two stage methodology will be introduced to argue why both handcrafted features and deep-encoding features with semantics is necessary for the grader. As shown in the figure x(number) (figure represents the processes after cleaning of the data), during the first stage after minimal cleaning of the data hand crafted features are calculated and concatenated with the combined scores of divisions (semantic score and prompt relevant scores) in second stage of the methodology.

A) Data cleaning

The data set used in the project is Automated student assessment price (ASAP). Essays in this dataset has tagged labels which as to be removed to make the data conducive for finding handcrafted features.

No	Features
1	Character's average and range of word lengths.
2	Characters average and range of sentence lengths in words
3	Essay length in words and characters
4	Prepositional and comma usage
5	word count for original essays
6	average amount of clauses per sentence
7	Mean clause length
8	Maximum number of clauses of a sentence in an essay

Table 3 features explained

B) Handcrafted features

Above in the table (number) are few handcrafted features used in [rp]. Some of the 8 essay types in ASAP data like essay set-2 are not only scored based on the writing applications but also language the language conventions which considers features such as spelling mistakes, punctuation errors, grammatical errors, paragraphing, word count etc. Generally, low scored essays contain less words and sentences

And some might contain very lengthy sentences which implies writer's lacking skills in writing concisely, features like vocab size i.e., the number of unique words in the essay helps us understand the writer's vocabulary level.

For grammar and spell check it only make sense to use some existing grammar error correction systems as they check for spelling errors, uncased letters, context, punctuation, word repetition etc. In the model languagetool[1] will be used as its one of the most accurate systems [3] and ginger it[2]. To collect handcrafted features and conduct error correction.

C) Sentence embedding/vectorization

It's not exaggeration to say that Google's Bidirectional Encoder Representations and transform (BERT) gave state of the art embedding results and proved itself to be better than word2vec in NLP tasks such as text classification [12]. Given the achievements of BERT, in this paper sentence embedding is performed on the stop words removed texts using $BERT_{base}$[2]. $BERT_{base}$ is a pretrained BERT model trained on 768 dimensions and 12 layers. "Pooled output" of the BERT result will be used to represent text into a low-dimensional embedding. **pooled output**: Is representation/embedding of CLS token passed through some more layers Bert pooler, linear/dense and activation function. It is recommended to use this pooled output as it contains contextualize information of whole sequence.

D) Semantic score

Deep semantic features are essential to analyse the semantic soundness and also for prompt relevancy checking in prompt dependant tasks. Sequential model is used with LSTM to map our data into a low-dimensional embedding and pass it to the dense layer for scoring the essays.

For every essay $e_x = \{s1, s2, \cdots, sm\}$, where s_k indicates the k th ($1 \leq k \leq m$, $s_k \in R^d$) embedded sentence in the essay and d = 768 means the length of sentence embedding. The encoding process of LSTM is described as follows [8]:

$$i = \sigma[W_i \,.\, s_t + U_I . h_{t-1} + b_i]\, c \quad - \ 1$$

$$f_t = \sigma\left[W_f \,.\, s_t + U_f . h_{t-1} + b_f\right] \quad - \ 2$$

$$C_t^{\sim} = \sigma[W_c . s_t + U_c . h_{t-1} + b_c] \quad - 3$$

$$C_t = it \,^{\circ} c_t^{\sim} + f_t {}^{\circ} c_{t-1} \quad - 4$$

$$0_t = \sigma[W_0 \,.\, S_t + U_0 \,.\, h_{t-1} + b_0] \quad - 5$$

$$h_t = 0_t \,^{\circ} tanh(c_t) \quad - 6$$

h_t - hidden state of sentence st. (W_x, U_x) where x=(I,f,c,o) are the weight matrices for the input, forget , candidate and output gates respectively. b stands for the bias vectors of the specific gates. sigmoid function= σ and ◦ means element-wise multiplication. Hence, for every essay, we will get the hidden state set H = {h_1, h_2, · · ·, h_m}. H_m is passed to dense layer to convert to scalar value. The values from dense layer output are projected in the range of 0-1 as ASAP dataset has different sets of essays with different scoring range mentioned in Table X

Prompt	Essay	Avg Length	Score Range
1	1783	350	2-12
2	1800	350	1-6
3	1726	150	0-3
4	1772	150	0-3
5	1805	150	0-4
6	1800	150	0-4
7	1569	150	0-30
8	723	650	0-60

Table 4: grading values

E) Prompt relevancy score

This score shows how relevant the essay is to the question/prompt. Just like semantic score same LSTM model is used. Prompt = {s_1 ……..,s_k} and essay={s_1 ……..,s_k} where s refers to sentence are combined into one set of sentences. Now the exact same process followed for semantic score will be followed here i.e., after LSTM hidden layer data is fed to dense layer which gives a scalar output scaled to 0-1 range and score 0 is received by essays that are irrelevant.

F) Training and evaluation metrics

It is known that essay sets ASAP dataset have different range of scoring so to be consistent all scores in each set in the training dataset are scaled down to 0-1 scale and will be re-scaled to original values during testing phase. The LSTM neural network used in our model is mono layered with 1024 as the size of hidden layer. It is known that the performance of multi layered and bidirectional LSTM model is subpar. A dropout proportion of 0.5 is set in order to avoid overfitting of the model

5.RESULTS DISCUSSION

Starting with implementation we will be using the pandas library to load the dataset from the folders. We shall remove some columns when reading the dataset. This dataset is free of trivial columns leaving only those columns that show cumulative scores. Find the handcrafted features by suing tools like ginger it, language etool , python etc. The handcrafted features are to be converted into a dataset mapping to the y labels of the essay dataset. Essay data will now be pre-processed and converted into list of vectors. The list of vectors will be fed to a deep learning model CNN+LSTM and DENSE layer, predict and find kappa score. Follow the same process for handcraft dataset. Concatenate prompts and their respective essays and perform the same task as last step to get the prompt relevant score. Concatenate all three scores using XGboost to get the final kappa scores. Print the spelling and grammar errors. Data Visualisation: Bar charts, heat maps, and other visualisations are used to describe the dataset's information.

Developing an AWE that works consistently for any kind of essay is vital. BERT proved itself to be the finest when it comes to contextual text embedding and the pretrained BERT base models vast enough. Handcrafted features definitely had their own significance. Provided the right set of features are chosen in the model with right tools to extract those features. The two-stage methodology is definitely a step up when compared to normal models giving us an unparalleled kappa score of 0.821 and is consistent across all types of essays in the data set. This proves that using handcrafted features combined with deep semantic features with BERT gives adequate results. Results shown in the table-3 are sourced from different papers using same dataset. EASE based on SVR and BLRR are used in [*"Flexible domain adaptation for automated essay scoring using correlated linear regression"*] to compute the results, CNN and CNN combined with LSTM are proven to having achieved great performance in many existing models [3].

```
array([[  0,   3,   7,   5,   0,   1,   0,   0,   0,   0],
       [  1,  41, 157,  91,  22,   5,   1,   2,   0,   0],
       [  0,   8,  67,  66,  27,   5,   0,   0,   0,   0],
       [  0,   3,  80, 130,  67,  32,  12,   2,   1,   0],
       [  0,   0,  23,  88,  45,  40,  11,   2,   1,   0],
       [  0,   1,  11,  41,  88, 107,  81,  29,   4,   0],
       [  0,   0,   0,   4,  30,  57,  38,  38,   4,   0],
       [  0,   0,   0,   2,  16,  36,  95, 120,  49,   0],
       [  0,   0,   0,   0,   0,   2,   6,  45,  38,   0],
       [  0,   1,   0,   0,   0,   1,   8,  27,  38,   0]], dtype=int64)
```

Fig. 5.1- Confusion matrix

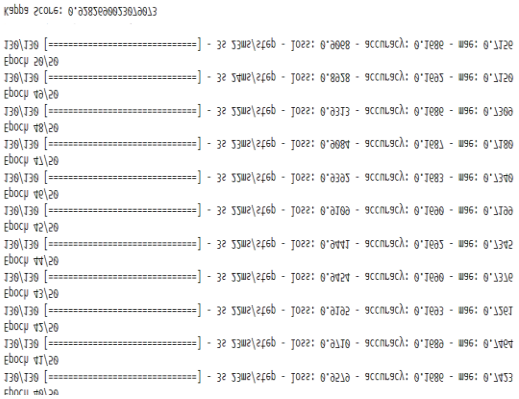

Fig. 5.2- final kappa score after 5 folds

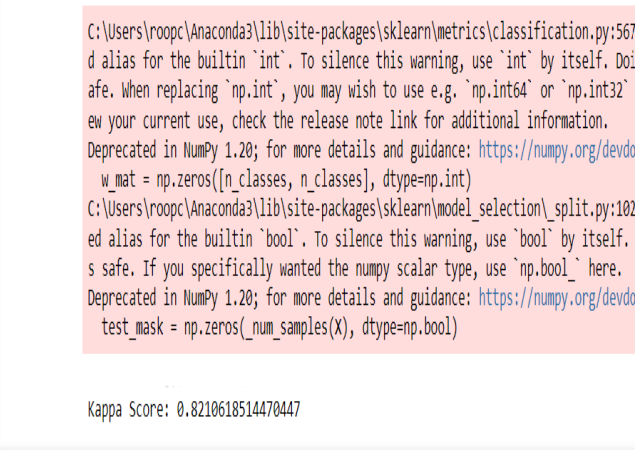

Fig. 5.3- Average kappa score after 5 folds

MODEL	SCORE
EASE(SVR)	0.699
EASE(BLRR)	0.705
CNN	0.726
LSTM	0.756
CNN+LSTM	0.761
TSM	0.821

Table 5.1- Comparison of different results from different papers

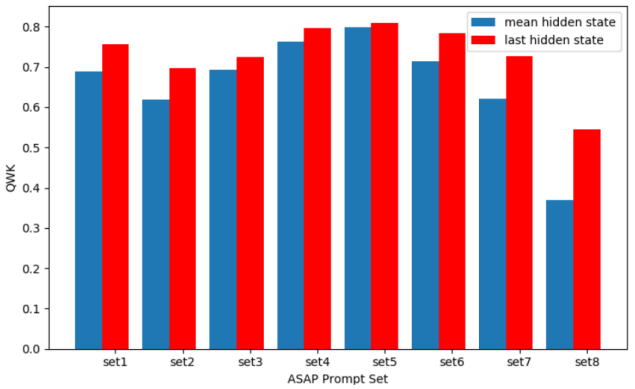

Fig5.4 - Set wise kappa score split [*8*]

6. CONCLUSION

We recommend the Two-Stage technique (TSM), which ensures the effectiveness and robustness of AES systems by utilising both hand-crafted and deep encoded features. The prompt relevant score PR and the semantic score Se in the initial step, two different sorts of scores called Pe are calculated. Both of these scores are based on LSTMT. Second, the handcrafted features score is determined. Third, these three scores are combined. The resulting data is input to a boosting tree model, which undergoes additional training and outputs a final score. The results of the studies demonstrate how effective TSM is on the ASAP dataset; our model outperforms numerous trustworthy baselines and performs better than average with a kappa score of about 0.82. In conclusion, both manually created and vector-encoded characteristics are used to provide our system its strength and capabilities.

REFERENCES

[1] [An In-Depth Comparison of 14 Spelling Correction Tools on a Common Benchmark](https://aclanthology.org/2020.lrec-1.228) (Näther, LREC 2020)

[2] T.Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” CoRR, vol. abs/1603.02754, 2016. [Online]. Available: http://arxiv.org/abs/1603.02754

[3] [A Neural Approach to Automated Essay Scoring](https://aclanthology.org/D16-1193) (Taghipour & Ng, EMNLP 2016)

[4] Alikaniotis, Dimitrios & Yannakoudakis, Helen & Rei, Marek. (2016). Automatic Text Scoring Using Neural Networks. 715-725. 10.18653/v1/P16-1068.

[5] Farag, Youmna & Yannakoudakis, Helen & Briscoe, Ted. (2018). Neural Automated Essay Scoring and Coherence Modeling for Adversarially Crafted Input. 263-271. 10.18653/v1/N18-1024.

[6] Jin, Cancan & He, Ben & Hui, Kai & Sun, Le. (2018). TDNN: A Two-stage Deep Neural Network for Prompt-independent Automated Essay Scoring. 10.18653/v1/P18-1100.

[7] Taghipour, Kaveh & Ng, Hwee. (2016). A Neural Approach to Automated Essay Scoring. 10.18653/v1/D16-1193.

[8] Liu, Jiawei & Xu, Yang. (2019). Automated Essay Scoring based on Two-Stage Learning.

[9] Yannakoudakis, Helen & Briscoe, Ted & Medlock, Ben. (2011). A New Dataset and Method for Automatically Grading ESOL Texts 180-189.

[10] S. Prabhu, K. Akhila and S. S, "A Hybrid Approach Towards Automated Essay Evaluation based on Bert and Feature Engineering," *2022 IEEE 7th International conference for Convergence in Technology (I2CT)*, 2022, pp. 1-4, doi: 10.1109/I2CT54291.2022.9824999.

[11] Z. Zhang and Y. Zhang, "Automated Writing Evaluation System: Tapping its Potential for Learner Engagement," in IEEE Engineering Management Review, vol. 46, no. 3, pp. 29-33, 1 thirdquarter,Sept. 2018, doi: 10.1109/EMR.2018.2866150.

[12] A. Alqahtani and A. Alsaif, "Automatic Evaluation for Arabic Essays: A Rule-Based System," 2019 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), 2019, pp. 1-7, doi: 10.1109/ISSPIT47144.2019.9001802.

[13] [Should You Fine-Tune BERT for Automated Essay Scoring?](https://aclanthology.org/2020.bea-1.15) (Mayfield & Black, BEA 2020)

[14] [On the Use of Bert for Automated Essay Scoring: Joint Learning of Multi-Scale Essay Representation](https://aclanthology.org/2022.naacl-main.249) (Wang et al., NAACL 2022)

[15] Y. Yang, L. Xia and Q. Zhao, "An Automated Grader for Chinese Essay Combining Shallow and Deep Semantic Attributes," in IEEE Access, vol. 7, pp. 176306-176316, 2019, doi: 10.1109/ACCESS.2019.2957582.

[16] Hongbo Chen, Jungang Xu, Ben He, Automated Essay Scoring by Capturing Relative Writing Quality, *The Computer Journal*, Volume 57, Issue 9, September 2014, Pages 1318–1330, https://doi.org/10.1093/comjnl/bxt117

[17] Burstein, Jill & Kukich, Karen & Wolff, Susanne & Lu, Chi & Chodorows, Martin & Braden-harderss, Lisa & Harrissss, Mary. (2002). Automated Scoring Using A Hybrid Feature Identification Technique. Vol. 1. 10.3115/980451.980879.

[18] Z. Wang, J. Liu and R. Dong, "Intelligent Auto-grading System," 2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS), 2018, pp. 430-435, doi: 10.1109/CCIS.2018.8691244.