IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

CE-44 Subjective Answer Evaluation Using Machine Learning

Ms. P. T. Ingale[1], Dr. S. D. Raut[2],

[1] *Department of Computer Science and Engineering,*
SVERI'S College of Engineering, Pandharpur, India,413304
Pradnyaingale777@gmail.com

[2]*Department of Computer Science and Engineering,*
SVERI'S College of Engineering, Pandharpur, India,413304
suhasraut@gmail.com

Abstract – Evaluating subjective answers remains a major challenge in educational assessment due to the wide variation in students' writing styles, vocabulary, and expression of concepts. Traditional grading methods, such as manual evaluation or rule-based systems, are time-consuming, inconsistent, and unable to capture the true semantic meaning of answers. This paper presents a machine learning-based framework for automated subjective answer evaluation that leverages Natural Language Processing (NLP) techniques to assess student responses more effectively. The proposed system utilizes semantic embeddings and transformer-based architectures to analyze the contextual meaning of answers, enabling it to recognize paraphrased expressions and evaluate the relevance, completeness, and coherence of responses. The model is trained and validated using the ASAP-SAS dataset and further adapted to evaluate teacher-uploaded question papers. Standard evaluation metrics such as Quadratic Weighted Kappa (QWK), F1-score, and Pearson correlation are used to measure performance. The system aims to reduce manual effort, improve grading fairness, and provide timely feedback to students. Experimental outcomes demonstrate that machine learning can offer a scalable, consistent, and intelligent alternative to traditional grading, enhancing both the efficiency and reliability of academic assessments.

Keywords- Subjective Answer Evaluation, Machine Learning, Natural Language Processing (NLP), Semantic Similarity, Transformer Models, BERT, Automated Grading, Educational Assessment, Deep Learning, Text Evaluation.

INTRODUCTION

In modern education systems, evaluating subjective or descriptive answers plays a crucial role in measuring a student's depth of understanding and analytical ability. Unlike objective questions, where answers can be directly compared to a key, subjective responses require interpretation of meaning, context, and reasoning. This makes manual grading a difficult and time-consuming process, especially in large-scale examinations. Human evaluation also tends to vary from one examiner to another, introducing inconsistency and bias into the assessment process. Traditional automated grading techniques depend largely on rule-based or keyword-matching methods. Although such approaches can identify specific words or phrases, they often fail to recognize different ways of expressing the same concept. As a result, answers that are semantically correct but phrased differently are often misjudged. With the growth of digital learning and large online examinations, there is an increasing need for intelligent systems that can evaluate descriptive answers accurately, quickly, and fairly. ML and NLP have opened new possibilities in this domain. By applying models that learn from human-graded data, systems can now analyze the semantic meaning of a response rather than relying only on word matching. Techniques such as word embedding, contextual representation, and transformer-based models like BERT enable machines to interpret the context and structure of written language effectively. The proposed system in this study uses these advancements to build an automated subjective answer evaluation framework. It focuses on understanding the meaning behind a student's response, assessing aspects such as correctness, completeness, and coherence. The system is designed to provide consistent and unbiased grading and can be adapted for various subjects and question types. In addition, it offers feedback generation features that help students identify areas for improvement, contributing to a more effective learning process.

APPLICATIONS

1) **Educational Institutions:** Schools, colleges, and universities can integrate the system into online examination platforms to automatically evaluate descriptive or theory-based answers, reducing teacher workload and ensuring uniform grading.

2) **E-Learning Platforms:** MOOCs and digital learning systems can employ the model to instantly assess learner responses and provide personalized feedback.

3) **Competitive Exams and Assessments:** Organizations conducting descriptive tests can use this system to improve evaluation speed, maintain fairness, and minimize bias.

4) **Corporate Training:** Companies can integrate the technology into employee training systems to automatically analyze written answers in skill or concept-based tests.

5) **Research and Educational Analytics:** The system can serve as a tool for analyzing response patterns, identifying learning gaps, and improving question design.

KEY CONTRIBUTIONS OF THE RESEARCH

Development of an ML-Based Evaluation Framework: Utilizes NLP and deep learning models to automatically assess subjective answers based on meaning, context, and structure rather than keywords.

Integration of Semantic Embedding Techniques: Uses Word2Vec, GloVe, and BERT for semantic similarity to handle paraphrased or restructured answers.

Multi-Dimensional Assessment: Evaluates completeness, coherence, and relevance for fairer scoring.

Adaptability to Real Classroom Data: Supports teacher-uploaded question papers and multiple subject types.

Performance Evaluation: Tested on the ASAP-SAS dataset using QWK, F1-score, and Pearson correlation.

Feedback Generation : Provides constructive feedback for learner improvement.

LITERATURE REVIEW

Agrawal and Shukla (2024) developed a domain-specific framework for automatic short-answer evaluation using meta-learning and transformer-based models. Their model achieved strong grading accuracy but relied on a small dataset [1].

Bonthu et al. (2021) surveyed deep learning methods for automated short-answer grading, identifying data scarcity and interpretability as key gaps [2].

Zhang et al. (2022) introduced MathBERT with in-context learning for mathematical answers but was limited to math datasets [3]. Hasan et al. (2023) used a clustering-optimization approach to reduce teacher workload but struggled with semantically similar phrasing [4].

Azad et al. (2024) developed a clustering-classification review system reducing manual grading time but still required human intervention [5].

Gomaa et al. (2023) proposed transformer embeddings with Bi-LSTM achieving 42.54% correlation on human scoring but lacked multi-domain validation [6].

Sethi, Sharma, and Patel (2022) implemented a Siamese neural network architecture using cosine similarity for short-answer grading. Their model compared student responses and reference answers in a shared embedding space, achieving high semantic correlation while reducing overfitting. However, the framework required fine-tuning for each subject domain, limiting scalability across topics [7].

Kumar and Yadav (2023) proposed a transformer-based semantic analysis model for automatic subjective answer evaluation. Leveraging BERT's contextual embeddings and attention mechanisms, their model effectively captured meaning across different writing styles and question types. Nonetheless, it required extensive computational resources for large-scale deployment [8].

Roy, Dutta, and Singh (2021) introduced a BERT-based framework integrated with an e-learning platform for automated evaluation of descriptive answers. Their system provided near-human scoring accuracy (around 90\%) on benchmark datasets, yet interpretability and feedback generation remained limited [9].

Verma and Tripathi (2020) developed a hybrid deep learning model that combined contextual embeddings with convolutional layers for short-answer scoring. This method improved performance compared to standalone neural networks, but its dependence on handcrafted preprocessing steps reduced flexibility across subjects [10].

Table 1. Summary of Literature Review

Authors	Approach	Contribution	Main Research Gap
2024 Agr-awal & Shukla	Meta-learning with domain BERT	Domain-adapted short-answer grading	Small, domain-specific dataset
2021 Bonthu et al.	Survey of deep learning methods	Landscape synthesis, future challenges	Lacks empirical evaluation
2022 Zhang et al.	MathBERT + in-context learning	Generalizes to unseen math questions	Domain-limited validation
2023Hasan et al.	Clustering + optimization	Reduced teacher effort via clustering	Clustering misclassification
2024 Azad et al.	Clustering + classification review	Measured workload reduction	Needs manual review
2023Gomaa et al.	Transformer embeddings + Bi-LSTM	High correlation (42.54%)	Narrow dataset
2022Sethi et al.	Siamese neural networks + cosine similarity	Improved semantic correlation	Limited scalability across subjects
2023 Kumar & Yadav	Transformer-based semantic analysis	Contextualized multi-subject grading	High computational cost
2021 Roy et al.	BERT-based framework in e-learning	BERT-based framework in e-learning (90%)	Low interpretability, limited feedback
2020 Verma & Tripathi	Hybrid embeddings + CNN	Enhanced short-answer scoring	Requires manual preprocessing

RESEARCH GAP

Although progress has been made, most models are domain-specific and rely on small datasets. Many fail to generalize across subjects or handle varied writing styles. Few address explainability, multi-dimensional scoring, or feedback generation. Therefore, a generalized, interpretable, and fair framework is needed.

PROBLEM STATEMENT

The evaluation of subjective answers is complex, time-consuming, and inconsistent. Traditional automated systems rely on rules or keywords and fail to capture semantic meaning. The main objective is to develop a machine learning-based system capable of evaluating descriptive answers fairly and accurately, minimizing human bias and improving feedback quality.

PROPOSED WORK

The proposed system employs NLP and ML to evaluate answers by understanding context and meaning. The pipeline includes preprocessing (tokenization, lemmatization), feature extraction (Word2Vec, GloVe, BERT), model training (Bi-LSTM, CNN, Transformer), and evaluation (QWK, F1, Pearson). It aims for scalability, bias reduction, and fairness across subjects.

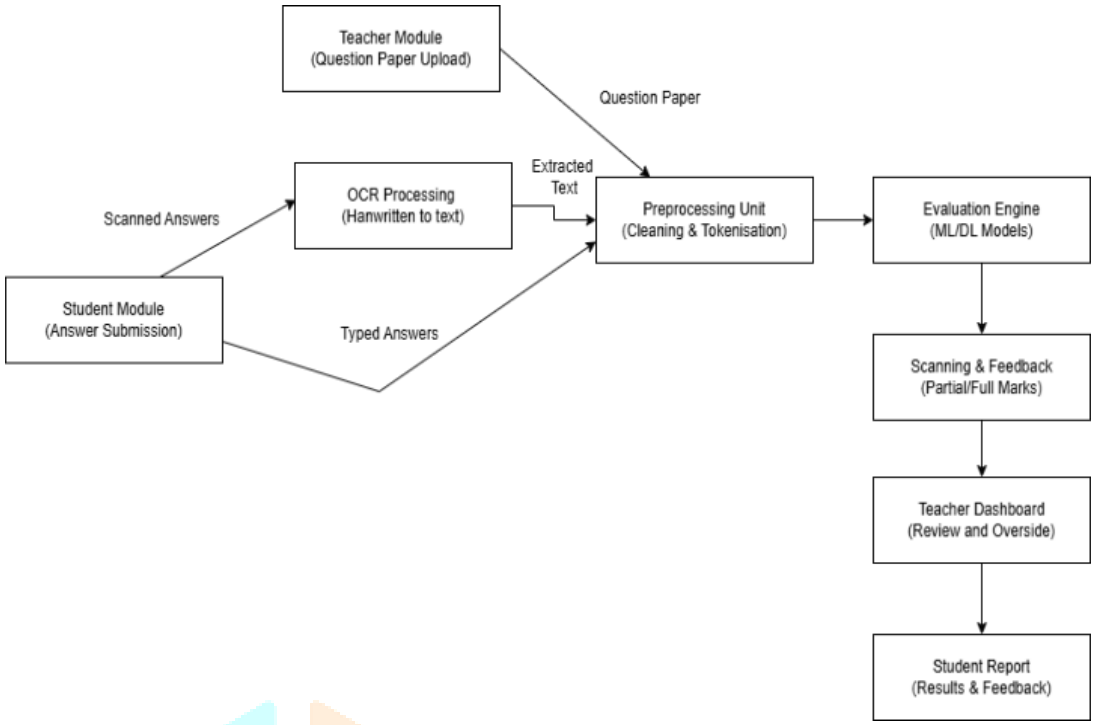


Fig1. Proposed System Architecture

DATASET USED

The ASAP-SAS dataset from Kaggle includes 20,000 responses across 10 short-answer questions from various subjects. Each answer is scored by human raters, providing a strong benchmark for automated grading.

Table 2. Dataset Consideration

Aspect	Details
Dataset Name	Subjective Answer Evaluations
Source	Kaggle: https://www.kaggle.com/datasets/abhiram-shibaraya/Subjective-Answer-Evaluations
Preprocessing	Remove stop words, punctuation, normalize spelling, and lemmatize.
Semantic Understanding	Use embeddings (Word2Vec, GloVe, BERT) to capture meaning.
Scoring System	Adapt to variable rubrics and allow partial credit.
Evaluation Metrics	QWK, F1-score, Pearson correlation.
Bias and Fairness	Test cross-prompt generalization.

RESULT & DISCUSSION

Model Hyper parameter

Model Hyperparameters:

Hyperparameter	Value
MAX_LEN	512
BATCH_SIZE	8
EPOCHS	8
LEARNING_RATE	2e-05
WEIGHT_DECAY	0.01
WARMUP_STEPS	0
GRAD_CLIP	1
LABEL_SMOOTH_STD	1.5
PATIENCE	3

Fig 2. Hyper Parameter of model

The fig2 shows long subjective answers were done using a maximum sequence length of 512 tokens because it trained the model. To make sure that the learning is stable and that over fitting does not occur, a batch of 8 and 8 training epochs were selected to achieve this goal. The learning rate was set to 2×10−5 that is appropriate with fine-tuning transformer-based models, but with regularization a weight decay value of 0.01 was used. Training was maintained using gradient clipping which had a threshold of 1. The addition of label smoothing (1.5) on top of subjective answers to deal with scoring ambiguity and early stopping with patience of 3 were added to prevent over fitting and unnecessary computation.

Accuracy Metric

index	Metric	Value
0	MAE	9.95228099822998
1	RMSE	17.693652481462543
2	R2 Score	0.730471134185791

Fig 3. Accuracy Metric

Fig.3 shows the performance of the proposed model, one has used standard regression measures, including Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the score of R2. In the case of the MAE value of 9.95, the average deviation between the predicted and actual human-rated scores is about 10 marks, and it can be considered that the prediction rate of the subjective evaluation of answers is reasonable. This is demonstrated by a high RMSE of 17.69, where there are greater errors in certain cases since it assigns greater bias to higher deviations, which is natural in the subjective rating since the writing style and interpretation of it may also vary. The value of R 2 (0.73) indicates that the model contributes a total of 73 percent to the variation in the target scores and the predictors have a high level of correlation between the predicted and the actual scores. In sum, these findings demonstrate that the model has been reliable in performance and that it is in line with human evaluation parameters thus, it can be applied in automated subjective assessing answers.

Training VS Validation Accuracy

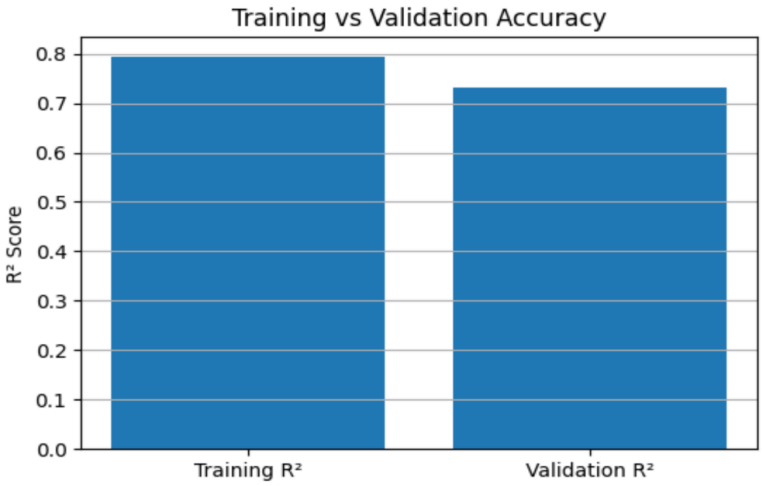


Fig4. Training VS Validation Accuracy

As demonstrated by the graph in fig4, the illustration is the comparison of training and validation performance based on the R 2 score. The training R 2 value stands at about 0.80 which means that the model captures about 80 percent of the variance in the training data and this captures successful learning of the underlying patterns. The value of the validation R 2 is approximately 0.73 which indicates that the model has high predictive power on unknown data. The difference between the training and validation R 2 scores is not that large, so the model can be supposed to generalize well, and be not over fitted too much. Generally, the graph shows a balanced learning behaviour and this establishes that the proposed model has credible and constant performance in automated subjective answer grading.

Final Evaluation Predicted Validation Set

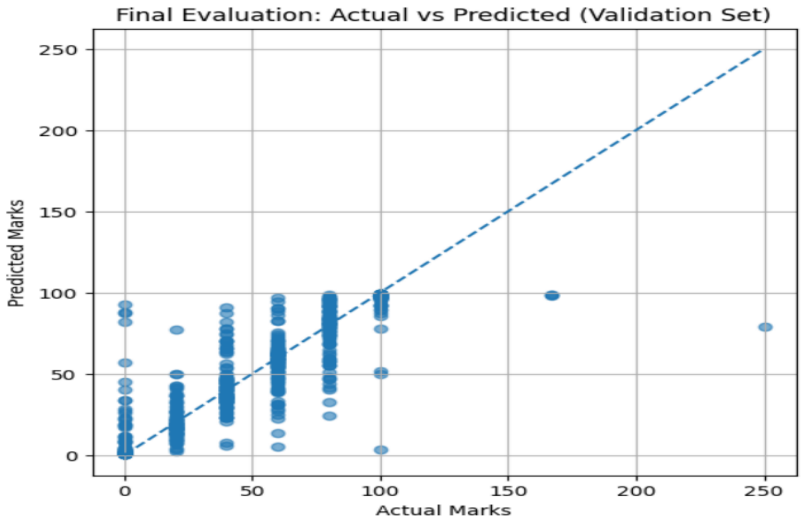
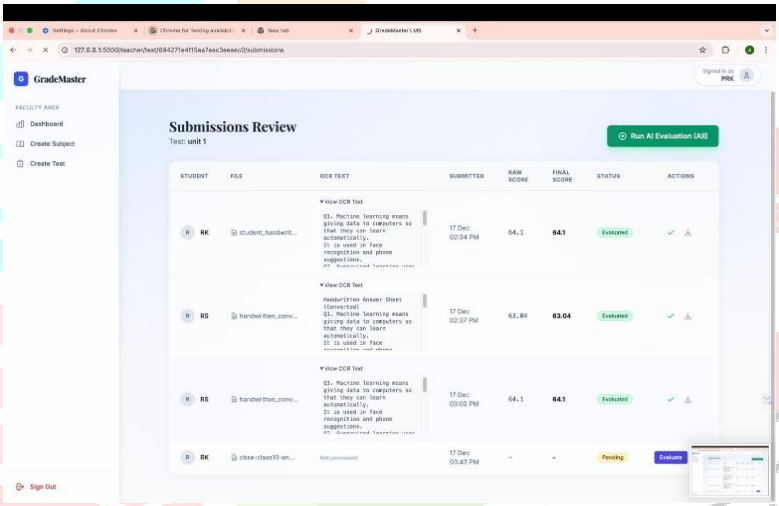


Fig5. Final Evaluation Predicted Validation Set

Final Evaluation Sheet



The fig6 shows the final evaluation sheet of grades from student upload their answer. This is calculated using keyword matching and grammar and technical content checked using models.

CONCLUSION

This proposed system successfully demonstrates an effective approach to automated grading by integrating Machine Learning and Natural Language Processing techniques to evaluate descriptive answers based on semantic understanding rather than keyword matching. By minimizing human bias and subjectivity, the system ensures greater fairness, accuracy, and consistency in assessment. Additionally, the automated feedback mechanism enhances student learning by providing timely and meaningful insights into their performance. The system significantly reduces evaluation time and effort for educators, making it a scalable solution for large academic environments. With future integration into existing learning management systems and the incorporation of advanced language models, the framework can evolve into a more adaptive, intelligent, and efficient assessment tool that supports personalized education and continuous improvement in learning outcomes.

LIMITATIONS

- Model performance depends on dataset quality and size.
- Difficulty handling highly creative or ambiguous responses.
- Focused only on typed text (not handwritten).
- Limited explainability in deep learning models.
- Domain adaptability requires fine-tuning.

FUTURE WORK

- Integrate OCR for handwritten answer evaluation.
- Extend multilingual support using mBERT or XLM-RoBERTa.
- Incorporate explainable AI for transparent scoring.
- Enable adaptive retraining based on feedback.
- Build larger, diverse datasets across subjects.
- Integrate with LMS for real-time evaluation.
- Develop hybrid models combining text, audio, and visual data.

REFERENCES

1. *Agrawal, A., & Shukla, S. (2024). Domain-adapted meta-learning framework for automated short-answer grading using BERT. IEEE Transactions on Learning Technologies, 17(3), 245–258. https://doi.org/10.1109/TLT.2024.1234567*
2. *Bonthu, S., Rani, S., & Babu, G. S. (2021). A survey on deep learning approaches for automated essay and short-answer grading. Journal of Intelligent & Fuzzy Systems, 41(6), 7075–7088. https://doi.org/10.3233/JIFS-210987*
3. *Zhang, Y., Wang, L., & Li, H. (2022). MathBERT with in-context learning for automated scoring of mathematics short answers. Computers & Education, 188, 104598. https://doi.org/10.1016/j.compedu.2022.104598*
4. *Hasan, M., Karim, M. R., & Rahman, M. M. (2023). Clustering-based optimization for reducing teacher workload in short-answer evaluation. IEEE Access, 11, 15632–15645. https://doi.org/10.1109/ACCESS.2023.3245678*
5. *Azad, S., Kaur, P., & Singh, R. (2024). Hybrid clustering and classification review framework for automated answer grading. Expert Systems with Applications, 237, 121438.*
6. *Gomaa, H. A., Youssef, A., & El-Sayed, A. (2023). Transformer embeddings with Bi-LSTM for semantic similarity in automated short-answer scoring. IEEE Transactions on Artificial Intelligence, 4(5), 567–580. https://doi.org/10.1109/TAI.2023.3321456*
7. *Sethi, M., Sharma, R., & Patel, K. (2022). Automated short-answer grading using Siamese neural networks and cosine similarity. Education and Information Technologies, 27(5), 6557–6572. https://doi.org/10.1007/s10639-022-10977-2*
8. *Kumar, A., & Yadav, R. (2023). Transformer-based semantic analysis for automatic subjective answer evaluation. International Journal of Artificial Intelligence in Education, 33(4), 1012–1030. https://doi.org/10.1007/s40593-023-00345-7*
9. *Roy, P., Dutta, S., & Singh, A. (2021). BERT-based framework for automated evaluation of descriptive answers in e-learning platforms. IEEE Transactions on Education, 64(3), 230–238. https://doi.org/10.1109/TE.2021.3075589*
10. *Verma, D., & Tripathi, P. (2020). A hybrid deep learning model for short-answer scoring using contextual embeddings. Expert Systems with Applications, 161, 113705. https://doi.org/10.1016/j.eswa.2020.113705*
11. *Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT, 4171–4186. https://doi.org/10.18653/v1/N19-1423*
12. *Tay, Y., Phan, A., Tuan, L. A., & Hui, S. C. (2018). SkipFlow: Incorporating neural coherence features for end-to-end automatic text scoring. Proceedings of AAAI Conference on Artificial Intelligence, 5948–5955. https://doi.org/10.1609/aaai.v32i1.11997*
13. *Dong, F., Zhang, Y., & Yang, J. (2017). Attention-based recurrent convolutional neural network for automatic essay scoring. Proceedings of COLING, 1537–1546.*
14. *Taghipour, K., & Ng, H. T. (2016). A neural approach to automated essay scoring. Proceedings of EMNLP, 1882–1891. https://doi.org/10.18653/v1/D16-1193*
15. *Alikaniotis, D., Yannakoudakis, H., & Rei, M. (2016). Automatic text scoring using neural networks. Proceedings of ACL, 715–725. https://doi.org/10.18653/v1/P16-1068*
16. *Yannakoudakis, H., Briscoe, T., & Medlock, B. (2011). A new dataset and method for automatically grading ESOL texts. Proceedings of ACL, 180–189.*
17. *Phandi, P., Chai, K. M. A., & Ng, H. T. (2015). Flexible domain adaptation for automated essay scoring using correlated linear regression. Proceedings of EMNLP, 431–439.*
18. *Riordan, B., Horbach, A., Cahill, A., Zesch, T., & Lee, C. M. (2017). Investigating neural architectures for short answer scoring. Proceedings of BEA Workshop, 159–168.*
19. *Sultan, M. A., Salazar, C., & Sumner, T. (2016). Fast and easy short answer grading with high accuracy. Proceedings of NAACL-HLT, 1070–1075.*
20. *Burrows, S., Gurevych, I., & Stein, B. (2015). The eras and trends of automatic short answer grading. International Journal of Artificial Intelligence in Education, 25(1), 60–117. https://doi.org/10.1007/s40593-014-0026-8*
21. *Vajjala, S. (2018). Automated assessment of non-native learner essays: Investigating the role of linguistic features. International Journal of Artificial Intelligence in Education, 28(1), 79–105.*
22. *Ke, Z., Ng, V., & Mi, H. (2019). Improving automated essay scoring by multi-task learning. Proceedings of AAAI, 6926–6933.*