AI in Education: Automated Essay Grading

FIRST REVIEW REPORT

Team Members

Sr. No	Reg. No	Student Name	School	Contact Number	E-mail Address
1	19BCE0241	Abhinav Gorantla	SCOPE	9791558808	abhinav.gorantla2019@vitstudent.ac.in
2	19BCE0243	Nishal Kulkarni	SCOPE	8237394877	nishal.kulkarni2019@vitstudent.ac.in

Slot - C1+TC1

Prepared For

Artificial Intelligence (CSE3013) – PROJECT COMPONENT

Submitted To

Prof. Geraldine Bessie Amali D

Submitted on 30-08-2021

School of Computer Science and Engineering

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

Problem Formulation

Problem Definition - Manual grading of essays is known to be a tedious process and poses several problems like grader's bias causing large variation in test scores and enormous wastage of time. There is a need for automation in this area and while auto-graders have existed in the past they have mostly involved basic keyword cross-checking which is not a sufficient metric to score essays.

Proposed Solution - Our solution is to build an automatic evaluation system which takes advantage of techniques from Natural Language Processing and Machine Learning to grade essays on the basis of features like tone, uniqueness of content, similarity between essays and grammatical complexity.

Goals - Develop and compare different models for effective essay grading and evaluating the impact of different metrics and features on the overall outcome(score in this case) for an essay.

Methodology - We will follow a quantitative approach towards meeting our goals, this primarily stems from the goal being achievable through comparison and analysis between performance of various classification algorithms.

Literature Survey

Our prior work involves a survey of 12 research articles relevant to automated essay evaluation.

Paper Title	Overview	Method	Critique
An automatic Arabic essay grading system based on text similarity Algorithms [1]	The research involves using various machine learning algorithms to grade essays written in arabic. The most promising results were given by using the N-gram technique.	It involves training solely on pre-graded Arabic test essays and then a comparison between string, corpus algorithms to achieve the best possible results.	1. Data set size is not large enough (only 210 essays) 2. End results might not be generalizable as they are more calibrated towards the semantics and morphology of the Arabic language
Automated essay grading using natural language processing and support vector machine [2]	Model is developed using the Education Testing Service's GRE Analytical Writing scoring guidelines. Uses SVM and NLP techniques to make a generalized essay grader.	Uses 8 sets each having ~1000 essays, extracting features and then comparing scores to the real scores given by human graders to measure error.	1. Does not involve similarity checking between multiple essays for the same test. 2. Essay topic is not considered when grading an essay.

Syntactic, semantic and sentiment analysis: The joint effect on automated essay evaluation [3]	The system incorporates checking similarity between semantics of the sentences using Graph based relationships on the contents of an essay to obtain unique features from it.	The algorithm uses 23 salient features not for just coherence checking on the surface level but to achieve a high predictive power. Results are measured using Quadratic Weighted Kappa.	1. Model lacks ontology based connections in the text. 2. Can benefit from using features based on centrality as is common in other graph networks.
Automated grading system using natural language processing [4]	We see a more primitive approach of traditional key-word cross checking here which involves maintaining SA (student answer) and CA(current answer) databases and then performing hit-miss percentile calculation.	The design of the algorithm is based on responses collected from teachers on the parameters which they follow during manual correction.	1. While the approach tries to incorporate additional features like synonyms comparison, most of it is still keyword cross-checking and better approach is possible (as proved by the papers already mentioned)
Memory-augmented neural model for automated grading [5]	The model is based on the idea that future work can be graded by using similar graded samples with score for each instruction in the rubric.	Student response grade is collected for every score in the rubric. The results are then compared with that of EASE (an open source AES system).	1. Model is tested only on a single dataset. 2. Additionally, the methods used for measuring relevance among these assignments is still rudimentary.
Automated essay grading using machine learning [6]	Performs a comparison of various machine learning techniques to find the best suited one for automated essay scoring. Regression trees achieve the best result.	With a data set size of 13000 essays this is the largest evaluation and comparison between Linear Regression, Regression Tree, Linear Discriminant Analysis, and Support Vector Machines.	1. Does not explore the structural or syntactical style but only goes uses the token content 2. Is outperformed by primitive methods which involve only word counts and grammatical error checking.
Machine learning approach for automatic short answer grading: A systematic review [7]	The research compares various Machine Learning approaches to Automatic Short Answer Grading (ASAG) systems.	A study to explore and understand the current state-of-the-art of ASAG with focus on works that used machine learning	1. Does a good overall comparison of several machine learning techniques for automatic grading. 2. More emphasis could

		approaches. Nature of datasets, NLP techniques and ML Algorithms used were taken into account while comparing various approaches to ASAG.	have been put on the features each of the models was using for evaluating an essay.
Automatic essay grading with probabilistic latent semantic analysis [8]	The model is based on Latent Semantic analysis which is a corpus based method used in information retrieval with vector space models. The relationship between meaning of a particular and the words in that particular essay was studied.	Probability that a word occurs in a particular context and other conditional probabilities were taken into account while assessing the essays. Comparison materials were taken either from a course book or other course materials.	1. Dataset could have been bigger. The one used in this study has about 300 essays.
An automatic short-answer grading model for semi-open-ended questions [9]	The model integrates domain-general and domain-specific information and also utilizes long-short-term-memory recurrent neural networks to learn the representation in the classifier .	Domain-specific information is integrated into the process of feature engineering. LSTM was used with softmax as the classifier. Multiple datasets were used to test the models.	1. Integrating domain-general and domain-specific information has significantly improved the performance of the neural network. 2. The largest disadvantage of this model is the requirement of many graded student answers for each question to train the model. This makes the model only useful for large-scale applications (millions of answers).
Automated essay scoring using generalized latent semantic analysis [10]	This Automatic Essay grading (AEG) system uses GLSA which makes n-gram by document matrix instead of word by document matrix.	All the essays being used for training the model were graded by multiple human graders and their average score was considered as the target score of that	1. Using GLSA instead of LSA significantly improved the performance of the AEG system.

		particular essay. Lingual errors were also taken into account while the model was grading an essay.	
Learning to grade short answer questions using semantic similarity measures and dependency graph alignments [11]	Graph alignment features along with lexical similarity measures have been combined with machine learning techniques to build an accurate automating grading system. A rudimentary deployment-graph method was employed to understand structural differences between sentences.	Similarity between dependency graphs is calculated for the expected answer and the submitted answer. In the second stage, the node similarity scores are used to weight the edges in a bipartite graph. In the final stage an overall grade is produced based on alignment scores found in the previous stage and some BOW similarity measures.	1. All the ranking systems used here had some kind of shortcomings. 2. The graph-alignment features introduced here weren't sufficient to act as a standalone grading system. But, it increased the performance of BOW features.
Automatic Short Answer Grading With SemSpace Sense Vectors and MaLSTM [12]	The proposed system is built on the MaLSTM network (Manhattan Long Short Term Memory). It is trained using the semantic similarity between reference answer and the student's response.	SemSpace algorithm was executed with WordNet data and 300-dimensional sense vectors were prepared. For the stop words, tokenization and lemmatization, Python NLTK library was used. Two dataset were used: Mohler dataset (available publicly) and CU-NLP dataset specifically generated for this research.	1. The use of SemSpace sense vectors has increased the accuracy of the current model by 30% when compared to the other auto grading systems which have used only LSTM.

Reference

[1] Shehab, A., Faroun, M., & Rashad, M. (2018). An automatic Arabic essay grading system based on text similarity Algorithms. *International Journal of Advanced Computer Science and Applications (IJACSA)*, *9*(3), 263-268.

[2] Suresh, A., & Jha, M. (2018). Automated essay grading using natural language processing and support vector machine. *International Journal of Computing and Technology*, *5*(2), 18-21.

[3] Janda, H. K., Pawar, A., Du, S., & Mago, V. (2019). Syntactic, semantic and sentiment analysis: The joint effect on automated essay evaluation. *IEEE Access*, *7*, 108486-108503.

[4] Rokade, A., Patil, B., Rajani, S., Revandkar, S., & Shedge, R. (2018, April). Automated grading system using natural language processing. In *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)* (pp. 1123-1127). IEEE.

[5] Zhao, S., Zhang, Y., Xiong, X., Botelho, A., & Heffernan, N. (2017, April). A memory-augmented neural model for automated grading. In *Proceedings of the fourth (2017) ACM conference on learning@ scale* (pp. 189-192).

[6] Song, S., & Zhao, J. (2013). Automated essay scoring using machine learning. *Stanford University*.

[7] Galhardi, L. B., & Brancher, J. D. (2018, November). Machine learning approach for automatic short answer grading: A systematic review. In *Ibero-american conference on artificial intelligence* (pp. 380-391). Springer, Cham.

[8] Kakkonen, T., Myller, N., Timonen, J., & Sutinen, E. (2005, June). Automatic essay grading with probabilistic latent semantic analysis. In *Proceedings of the second workshop on Building Educational Applications Using NLP* (pp. 29-36).

[9] Zhang, L., Huang, Y., Yang, X., Yu, S., & Zhuang, F. (2019). An automatic short-answer grading model for semi-open-ended questions. Interactive learning environments, 1-14.

[10] Islam, M. M., & Hoque, A. L. (2010, December). Automated essay scoring using generalized latent semantic analysis. In *2010 13th International Conference on Computer and Information Technology (ICCIT)* (pp. 358-363). IEEE.

[11] Mohler, M., Bunescu, R., & Mihalcea, R. (2011, June). Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies* (pp. 752-762).

[12] Tulu, C. N., Ozkaya, O., & Orhan, U. (2021). Automatic Short Answer Grading With SemSpace Sense Vectors and MaLSTM. *IEEE Access*, *9*, 19270-19280.