AUTOMATED ESSAY SCORING SYSTEM

[1]Sirisha Gummidi
Information Technology
Shri Vishnu Engineering College For Women (JNTUK)
Bhimavaram, India
santhisiri1987@gmail.com

[2]Bhagya Sai Deepika Golisetti
Information Technology
Shri Vishnu Engineering College For Women (JNTUK)
Bhimavaram, India
golisettideepika02@gmail.com

[3]Jyoshika Gunduboina
Information Technology
Shri Vishnu Engineering College For Women (JNTUK)
Bhimavaram, India
jyoshikagunduboina@gmail.com

[4]Isfehaani Sardaar Banu
Information Technology
Shri Vishnu Engineering College For Women (JNTUK)
Bhimavaram, India
sardaarbanu@gmail.com

[5]Dr.G.Ratnakanth (Professor)
Information Technology
Shri Vishnu Engineering College For Women (JNTUK)
Bhimavarm, India
ratnakanth@svecw.edu.in

***Abstract*— This study introduces an Automated Essay Scoring System (AESS) that is designed to evaluate and give feedback on written essays. The system uses natural language processing (NLP) techniques, including machine learning algorithms, like Random Forest . Recurrent neural networks (RNN) to analyze the content, structure and coherence of essays. By training on sets of human scored essays the AESS shows high accuracy in grading essays on various subjects and styles. Moreover it provides comments and suggestions for improvement to students to enhance their learning experience. Evaluation outcomes suggest that the AESS performs comparably to graders in essay scoring while offering benefits such as scalability, consistency and efficiency. This study contributes to the progress of automated assessment tools in education serving as an asset, for both educators and students.**

Keywords; Automated Essay Scoring System (AESS) Natural Language Processing (NLP) Machine Learning (ML) Feedback, Evaluation.

1. INTRODUCTION

Manually evaluating essays is time-consuming. Human graders may unintentionally exhibit bias when assessing essays[1]. The use of an unbiased training dataset can prevent inefficiencies in grading and inconsistency in feedback when utilizing automated essay scoring systems[2]. Consequently, there is a growing trend in the development and application of automated essay evaluation systems.

Since its inception, the fundamental process for Automated Essay Scoring (AES) has involved commencing with a training set of essays that undergo meticulous manual scoring[3]. The program assesses surface features of the text in each essay, including total word count, the presence of subordinate clauses, or the ratio of uppercase to lowercase letters—measurable quantities that don't require human interpretation. Subsequently, it formulates a mathematical model linking these features to the assigned scores of the essays. This established model is then employed to compute scores for new essays.

Recently, a mathematical model of this nature was developed by Isaac Persing and Vincent Ng.[4]It assesses essays not only based on specified features but also considers the strength of the argument. The evaluation encompasses various aspects, including the author's level of agreement and the supporting reasons, adherence to the prompt's topic, identification of argument components (major claim, claim, premise), identification of errors in the arguments, and cohesion among other features. Unlike the previously mentioned models, this particular model closely emulates human judgment when grading essays. The increasing prevalence of deep neural networks has led to the adoption of deep learning methods for automated essay scoring, consistently achieving higher and sometimes surpassing levels of agreement seen among human graders[5].

The diverse Automated Essay Scoring (AES) programs vary in the specific surface features they measure, the size of the required training set, and notably, the mathematical modeling technique employed. Initial efforts relied on linear regression, while contemporary systems may utilize linear regression or other machine learning methods, often complemented by additional statistical techniques such as latent semantic analysis[6] and Bayesian inference.[7]

Researchers have delved into the cross-domain aspect of automated essay scoring, employing various machine learning models to investigate its nuances and challenges. In this context, models are trained on essays composed for one prompt (topic) and then tested on essays written for a different prompt. Successful approaches in the cross-domain scenario often rely on deep neural networks [8] or models that combine deep and shallow features[9].

Automated essay scoring involves the computerized evaluation of essays, with grading models being developed through the analysis of essay datasets scored by different human graders[10]. Automated Essay Scoring (AES) represents a groundbreaking application of machine learning in the realm of education, seeking to automate the assessment of written essays. This pioneering system utilizes sophisticated natural language processing (NLP) and machine learning algorithms to evaluate and score essays, closely emulating the process of human grading. Nevertheless, offering targeted feedback to every student on numerous drafts of each essay throughout the school year poses a challenge, even for the most dedicated teachers[11].

Automated essay scoring allows students to engage in repetitive practice by taking tests and composing essays to enhance the quality of their responses. English proficiency examinations like GRE and TOEFL incorporate the e-rater (Writing evaluation) automated writing evaluation engine. The scores generated in these tests typically represent the combined average of the automated score and a human grader's assessment. The e-rater engine utilizes various features related to writing quality, including but not limited to grammar errors, usage, mechanics, style, discourse structure, sentence variety, source utilization, and discourse coherence quality[10].Research on writing evaluation and implementation commenced several decades ago, persisting and evolving towards more advanced automated evaluation systems.

The debate article authored by (Hearst, 2000)[12] presents the research work on essay grading or writing evaluation. It elucidates the progression of automated evaluation tools, tracing the development from PEG Writer's workbench to Short-answer scoring systems within the timeframe of 1960 to 2000. By 2000, some of the operational automated evaluation systems included PEG, e-rater, Latent Analysis, and Criterion.

(Burstein, Kukich, Wolfe & Chodorow, 1998)[13] constructed an electronic essay rater incorporating features such as discourse marking, syntactic information, and topical content. In their study, they compared two content vectors to predict scores, emphasizing both essay content and essay argument content. The electronic essay rater demonstrated an average accuracy of 82% when comparing argument content scores with human raters and 69% when comparing essay content with human raters. Notably, incorporating the discourse marking feature led to an impressive 87%-94% agreement between e-rater and human raters across 15 sets of essay responses.

(Crossley et al., 2016)[14] The discussion centers on achieving automatic essay quality assessment by integrating Natural Language Processing (NLP) and machine learning approaches to evaluate text features. Additionally, it explores assessing individual differences in writers through the collection of information from standardized test scores and survey results.

ReaderBench [15] is an open-source framework that functions as an automated text analysis tool, computing indices associated with linguistic and rhetorical features within the text. In a test involving 108 university students' essays, the framework demonstrated a 32.4% variance in vocabulary scores. Particularly noteworthy is its improved performance when applied to essays with multiple paragraphs.

Neural network models have found application in automated essay scoring. For instance, in their work, Fei et al. (2017) [16] utilized recurrent and convolutional neural networks to model input essays, determining grades based on a single vector representation of the essay.

(Woods et al., 2017)[17] elucidates the significance of acquiring effective writing skills in secondary education, which has subsequently led to the development of automated essay scoring. In their work, Woods et al. (2017) [17] explored an ordinal essay scoring model for generating feedback based on a rubric, employing predictive realistic essay variants. Writing Mentor TM, as an add-on, is specifically crafted to offer feedback to struggling writers with the aim of enhancing their writing skills. It leverages Natural Language Processing (NLP) techniques and resources to generate feedback, encompassing various writing sub-constructs. The tool has demonstrated positive results from users, both in terms of usability and its potential impact on improving their writing.[18].

The project on Automated Essay Scoring system signifies a substantial advancement in educational technology, presenting a practical solution to the complexities associated with manual essay grading. By incorporating machine learning and Natural Language Processing (NLP), the system strives to improve the evaluation process, offering educators a valuable tool for streamlined and dependable essay assessment. Indeed, certain researchers have reported that their Automated Essay Scoring (AES) systems can outperform human grading. Page, for instance, made this assertion for PEG back in 1994.[19]

2. LITERATURE REVIEW

Automated essay grading systems utilizing deep learning and deployed through Streamlit have attracted significant interest in recent times for their potential to transform the assessment process in educational environments.

Several research studies have investigated the effectiveness of automated essay grading systems using deep learning techniques like convolutional neural networks (CNNs) and recurrent neural networks (RNNs) in evaluating essay quality. These systems often yield similar results to human graders in terms of accuracy and dependability, presenting benefits in scalability and efficiency. Additionally, the incorporation of Streamlit aids in creating user-friendly interfaces, improving accessibility and usability for educators and students.

The existing literature emphasizes the ability of automated essay grading systems to simplify the grading process, enhance consistency, and offer prompt feedback to students. However, shortcomings such as the reliance on predetermined features, the failure to capture subtle writing nuances, and concerns regarding fairness and bias highlight the necessity for ongoing enhancement and assessment of these systems.

Despite progress in automated essay grading using deep learning and Streamlit, discrepancies and gaps persist in the literature. These include disparities in dataset composition and size, inconsistencies in evaluation criteria, and limited applicability across various writing themes and fields.

A consensus from the literature points towards the potential of deep learning models, especially CNNs and RNNs, in bolstering the accuracy and effectiveness of automated essay grading systems. Furthermore, there is a growing inclination towards incorporating Streamlit for developing user-friendly interfaces to address usability challenges linked with traditional automated essay grading platforms.

The existing literature shapes our research inquiries by underscoring the need to tackle challenges such as dataset diversity, model interpretability, and ethical considerations in the development of automated essay grading systems.

Based on the gaps and unanswered queries identified in the literature, potential areas for further exploration include:

- Exploration of new deep learning architectures for automated essay grading with enhanced interpretability and equity.
- Investigation of diverse datasets covering multiple writing prompts and language variations.

- Study of user perceptions and experiences with Streamlit-based automated essay grading interfaces to guide interface design and usability improvement

3. RESEARCH DESIGN AND METHODOLOGY

3.1 Data Set Description

There are eight sets of essays, each generated from a prompt. The essays vary in length, ranging from 150 to 550 words. Responses were written by students in grades 7 to 10, and each essay was hand-graded and double-scored. Each data set has unique characteristics designed to push the limits of your scoring engine's capabilities.

The training data is available in three formats: a tab-separated value (TSV) file. Each file includes 28 columns:

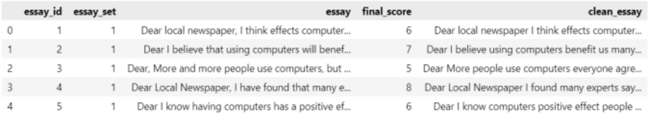
- essay_id: Unique identifier for each student essay
- essay_set: Indicates the set number (1-8) for each essay
- essay: The student's response in ASCII text
- rater1_domain1: Score from Rater 1's domain 1
- rater2_domain1: Score from Rater 2's domain 1
- rater3_domain1: Score from Rater 3's domain 1 (only present for some essays in set 8)
- domain1_score: Final score resolved between the raters
- rater1_domain2: Score from Rater 1's domain 2 (only present for essays in set 2)
- rater2_domain2: Score from Rater 2's domain 2 (only present for essays in set 2)
- domain2_score: Final score resolved between the raters for domain 2 (only present for essays in set 2)
- rater1_trait1 score - rater3_trait6 score: Trait scores for sets 7-8
- domain1_predictionid: This is a unique identifier corresponding to the predicted score for domain 1
- domain2_predictionid: Unique prediction_id corresponding to the predicted score for domain 2 (only present for essays in set 2)
- prediction_weight: Identifies the weight of the prediction in the final score calculation. For essay set 2, where two domains are scored, the weight is 0.5 to ensure equal contribution from each essay. For other essay sets, the weight is 1.0.
- predicted_score: it depicts the anticipated score of an essay.

3.2 Data Pre-Processing

The initial phase of our project involved standardizing the raw data through a comprehensive preprocessing pipeline. This included handling missing values by filling them appropriately and selecting pertinent features from the dataset. By addressing these initial data quality issues, we aimed to create a more robust foundation for subsequent analysis.

As a crucial step in understanding the distribution of our data, we employed graphical methods to assess skewness. Recognizing the importance of a balanced and normalized dataset for effective model training, we applied normalization techniques to mitigate any skewness observed. This normalization process ensures that the data adheres to a standardized scale, preventing any biases that may arise from varying scales across features.

Subsequently, we focused on refining the textual content of the essays to streamline the training process and enhance accuracy. This involved a meticulous cleaning process, wherein unnecessary symbols, stop words, and punctuation were removed. The goal was to distill the essays to their essential content, minimizing noise and irrelevant information.



	essay_id	essay_set	essay	final_score	clean_essay
0	1	1	Dear local newspaper, I think effects computer...	6	Dear local newspaper I think effects computer...
1	2	1	Dear I believe that using computers will benef...	7	Dear I believe using computers benefit us many...
2	3	1	Dear, More and more people use computers, but ...	5	Dear More people use computers everyone agre...
3	4	1	Dear Local Newspaper, I have found that many e...	8	Dear Local Newspaper I found many experts say...
4	5	1	Dear I know having computers has a positive ef...	6	Dear I know computers positive effect people ...

Fig.1. Preprocessed Data Set

In our pursuit of improved accuracy, we decided to augment our feature set by incorporating additional linguistic and structural characteristics of the essays. This included features such as sentence count, word count, character count, and average word length. These features provide a more nuanced representation of the essays, capturing both macroscopic and microscopic aspects of their composition.

Moreover, we delved into linguistic analysis by exploring the frequencies of different parts of speech—nouns, verbs, adjectives, and adverbs. Employing parts of speech tagging, we gained insights into the composition and syntactical structure of the essays, enriching our feature set with linguistic nuances.

To further enhance our linguistic analysis, we implemented a strategy to identify misspellings in the essays. This involved comparing the essays to a predefined corpus, enabling the detection and correction of misspelled words. This step aimed to improve the overall quality of the textual content and contribute to a more accurate assessment.

The final phase of our approach involved the application of various machine learning algorithms to the preprocessed and enriched dataset. The choice of algorithms was guided by the specific characteristics of our data and the nature of the automated essay scoring task. Details of these algorithms and their performance are outlined in the subsequent section, providing a comprehensive overview of our modeling strategy.

3.3 Feature Extraction

To prepare our dataset for the application of machine learning algorithms, a crucial step involves converting the textual content of essays into a numeric format. Machine learning algorithms inherently operate on numeric data, and for this purpose, we employed a technique known as CountVectorizer. This process involves tokenizing a collection of text documents, breaking them down into individual words (or tokens), and encoding them into numeric vectors. The resulting vectors have a length equivalent to the entire vocabulary, with each element representing the count of occurrences of a specific word in the corresponding document. This transformation is pivotal as it enables the utilization of machine learning algorithms on our essay dataset.

Initially, we subjected the dataset to machine learning algorithms such as linear regression, Support Vector Regression (SVR), and Random Forest without incorporating additional features that were highlighted during the preprocessing stage. Unfortunately, the initial results were unsatisfactory, evident from the high mean squared error across all the aforementioned algorithms. Recognizing the need for improvement, we introduced extra features extracted during the preprocessing phase. Following this enhancement, we reapplied the CountVectorizer to the modified dataset and reran the same three algorithms.

The outcomes demonstrated a significant enhancement in the performance of all three algorithms, with particular emphasis on Random Forest. The mean squared error, a metric reflecting the accuracy of predictions, witnessed a drastic reduction. This improvement underscores the importance of feature engineering and thoughtful preprocessing in enhancing the efficacy of machine learning models. The refined dataset, enriched with additional features and transformed through CountVectorizer, proved to be instrumental in achieving more accurate and reliable predictive modeling results.

3.4 Applying Machine Learning Algorithm

To prepare our dataset for the application of machine learning algorithms, a crucial step involves converting the textual content of essays into a numeric format. Machine learning algorithms inherently operate on numeric data, and for this purpose, we employed a technique known as CountVectorizer. This process involves tokenizing a collection of text documents, breaking them down into individual words (or tokens), and encoding them into numeric vectors. The resulting vectors have a length equivalent to the entire vocabulary, with each element representing the count of occurrences of a specific word in the corresponding document. This transformation is pivotal as it enables the utilization of machine learning algorithms on our essay dataset.

Initially, we subjected the dataset to machine learning algorithms such as linear regression, Support Vector Regression (SVR), and Random Forest without incorporating additional features that were highlighted during the preprocessing stage. Unfortunately, the initial results were unsatisfactory, evident from the high mean squared error across all the aforementioned algorithms. Recognizing the need for improvement, we introduced extra features extracted during the preprocessing phase. Following this enhancement, we reapplied the CountVectorizer to the modified dataset and reran the same three algorithms.

The outcomes demonstrated a significant enhancement in the performance of all three algorithms, with particular emphasis on Random Forest. The mean squared error, a metric reflecting the accuracy of predictions, witnessed a drastic reduction. This improvement underscores the importance of feature engineering and thoughtful preprocessing in enhancing the efficacy of machine learning models. The refined dataset, enriched with additional features and transformed through CountVectorizer, proved to be instrumental in achieving more accurate and reliable predictive modeling results.

3.5 Applying Neural Networks

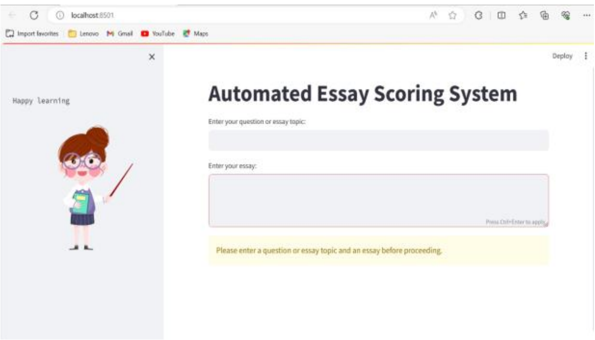
In the context of neural networks, preprocessing steps differ significantly from those employed in traditional machine learning algorithms. For neural network models, the training data undergoes a unique preprocessing sequence tailored to the utilization of an Embedding Layer, specifically Word2Vec. Word2Vec is a shallow, two-layer neural network designed to reconstruct the linguistic contexts of words. Its objective is to map words from a large corpus into a vector space, typically spanning several hundred dimensions. In this vector space, each unique word in the corpus is assigned a corresponding vector. The arrangement of these word vectors is such that words sharing common contexts in the corpus are positioned in close proximity to each other, capturing semantic relationships.

The features generated by Word2Vec are then fed into a Long Short-Term Memory (LSTM) network. LSTM is a type of recurrent neural network (RNN) that excels in learning dependencies and patterns in sequential data. It has the ability to discern which data in a sequence is essential to retain and which can be discarded. This capability is particularly advantageous for processing essays, allowing the model to capture the intricate relationships between words and sentences.

The final stage of the neural network architecture involves a Dense layer with an output of 1, serving as the predictor for essay scores. This layer uses the learned features from Word2Vec and LSTM to predict the overall score of each essay. The entire architecture is geared towards leveraging the power of neural networks to understand and represent the complex structures and relationships within textual data. By incorporating Word2Vec and LSTM, the model gains the ability to capture nuanced contextual information, making it well-suited for tasks such as automated essay scoring where understanding the sequence and context of words is crucial.

3.6 Integration using Streamlit

Integrating an Automated Essay Scoring (AES) project with Streamlit, a web framework for machine learning applications, presents numerous advantages that revolutionize the educational assessment landscape. By merging machine learning algorithms with a user-friendly web interface, this integration offers a seamless and intuitive experience for all users as it simplifies the submission process and facilitates instant feedback delivery. This streamlining of assessment procedures not only saves time but also enhances the overall efficiency of the learning process.



Automated Essay Scoring System

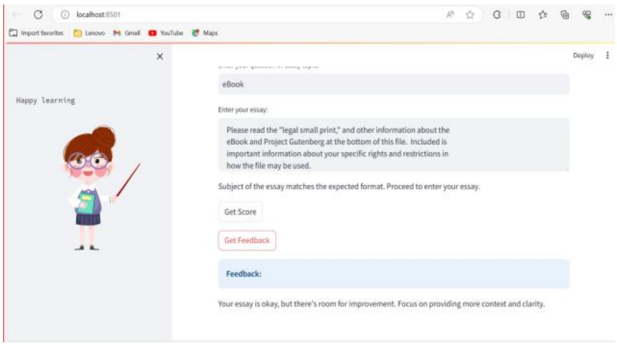
Enter your question or essay topic:

Enter your essay:

Please enter a question or essay topic and an essay before proceeding.

Fig.2. Integrated AES Screen with Streamlit

Moreover, the real-time nature of interaction enabled by Streamlit fosters swift and effective learning outcomes. Students receive immediate evaluations, enabling them to identify areas for improvement promptly.



eBook

Enter your essay:

Please read the "legal small print," and other information about the eBook and Project Gutenberg at the bottom of this file. Included is important information about your specific rights and restrictions in how the file may be used.

Subject of the essay matches the expected format. Proceed to enter your essay.

Feedback:

Your essay is okay, but there's room for improvement. Focus on providing more context and clarity.

Fig.3. Evaluating and providing feedback

Furthermore, Streamlit's visualization capabilities empower users to gain insights into performance analytics and trends. This analytical depth facilitates continuous improvement, ensuring that educational practices remain dynamic and responsive to evolving student needs.

In essence, the integration of AES with Streamlit not only automates essay grading but also transforms the entire educational assessment experience. By making assessment more interactive, accessible, and conducive to continuous improvement, this integration holds the potential to revolutionize educational practices and enhance learning outcomes for students across diverse settings.

4 . CONCLUSION

The Automated Scoring System represents a ground breaking advancement in evaluation methodologies, integrating machine learning and natural language processing (NLP) technologies. This project is seamlessly integrated with the web using Streamlit, enhancing accessibility and user interaction.

The significance of this system lies in its transformative impact on scoring procedures, showcasing the immense potential of cutting-edge technologies. Our experiment demonstrates the system's proficiency in accurately and efficiently assessing textual data across diverse domains, such as education and employment, surpassing the capabilities of traditional manual methods. The automation of scoring processes has yielded substantial improvements in terms of consistency, objectivity, and scalability.

The core objectives of our study were to mitigate subjectivity and bias while enhancing the reliability and efficiency of scoring methods. The outcomes of this research contribute substantially to the existing knowledge in assessment and evaluation procedures, particularly considering the rapidly evolving landscape of technology.

Despite facing challenges like data scarcity and algorithmic complexity, our study has paved the way for further exploration and refinement of automated scoring systems. Future endeavors should focus on expanding the understanding of NLP and machine learning applications in scoring procedures to maximize their effectiveness and applicability.

In conclusion, our study underscores the importance of embracing technological innovation to meet the evolving demands of assessment and evaluation processes. By incorporating cutting-edge approaches, such as NLP and machine learning, we can enhance the efficacy, efficiency, and fairness of scoring procedures.

5. FUTURE WORK

In considering the future development of our automated essay scoring project, several avenues for enhancement present themselves. Firstly, the system's feedback mechanism could evolve to offer fine-grained insights, breaking down the evaluation into specific aspects like grammar, structure, and citation usage for more targeted improvement suggestions. Additionally, the project could be extended to classify essays based on content types, such as Expository, Descriptive, Narrative, Compare-&-contrast, or Persuasive/argumentative, providing a deeper understanding of students' proficiency in various writing styles. Another area of improvement involves pinpointing specific paragraphs or sentences that may require attention, offering students a more granular understanding of their writing strengths and weaknesses. Expanding beyond English, the project could be adapted to support automated essay scoring in other languages, broadening its global applicability. Further advancements could include adaptive learning paths, personalized feedback based on individual progress, and the integration of cutting-edge Natural Language Processing techniques for more nuanced evaluations. Moreover, fostering collaboration through peer review within the automated system and tailoring scoring criteria to specific subjects could simulate real-world writing environments and ensure a more specialized assessment aligned with academic contexts. As the project continues to evolve, these considerations aim to enhance the system's comprehensiveness, adaptability, and accessibility in assessing and refining students' writing skills.

6. ACKNOWLEDGEMENT

We extend our sincere gratitude to Dr.D.V.Naga Raju(Head of the Department, Information Technology)for their invaluable guidance, support, and encouragement throughout this research endeavor. Their expertise and insights have been instrumental in shaping this study.

We express our heartfelt appreciation to Dr.G.Ratnakanth (Professor, Information Technology) for their unwavering assistance, patience, and constructive feedback, which significantly contributed to the refinement of this research.

We are deeply thankful to Shri Vishnu Engineering College For Women, Bhimavaram for providing the necessary resources and facilities essential for the completion of this project. Their support has been indispensable in facilitating our research efforts and also granting access to research materials and resources essential for conducting this investigation.

7.REFERENCES

[1] Zupanc, K., & Bosnić, Z. (2018). Increasing accuracy of automated essay grading by
grouping similar graders. Proceedings of the 8th International Conference on Web
Intelligence, Mining and Semantics - WIMS 18. doi:10.1145/3227609.3227645

[2] Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to
computer programmer as woman is to homemaker? debiasing word embeddings.
In Advances in neural information processing systems (pp. 4349-4357).

[3] Keith, Timothy Z. (2003), p. 149.

[4] Persing, Isaac, and Vincent Ng (2015). "Modeling Argument Strength in Student Essays", pp. 543-552. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Retrieved 2015-10-22.

[5] Yang, Ruosong; Cao, Jiannong; Wen, Zhiyuan; Wu, Youzheng; He, Xiaodong (2020). "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 2020. Online: Association for Computational Linguistics: 1560–1569. doi:10.18653/v1/2020.findings-emnlp.141. S2CID 226299478

[6] Bennett, Randy Elliot, and Anat Ben-Simon (2005), p. 7.

[7] Elliot, Scott (2003). "Intellimetric TM: From Here to Validity", p. 75. In Shermis, Mark D., and Jill Burstein, eds., Automated Essay Scoring: A Cross-Disciplinary Perspective. Lawrence Erlbaum Associates, Mahwah, New Jersey, ISBN 0805839739

[8] Cao, Yue; Jin, Hanqi; Wan, Xiaojun; Yu, Zhiwei (25 July 2020). "Domain-Adaptive Neural Automated Essay Scoring". Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '20. New York, NY, USA: Association for Computing Machinery. pp. 1011–1020. doi:10.1145/3397271.3401037. ISBN 978-1-4503-8016-4. S2CID 220730151

[9] Cozma, Mădălina; Butnaru, Andrei; Ionescu, Radu Tudor (2018). "Automated essay scoring with string kernels and word embeddings". Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Melbourne, Australia: Association for Computational Linguistics: 503–509. arXiv:1804.07954. doi:10.18653/v1/P18-2080. S2CID 5070986

[10] Shankar, R. S., & Ravibabu, D. (2018). Digital Report Grading Using NLP Feature
Selection. Soft Computing in Data Analytics Advances in Intelligent Systems and
Computing,615-623. doi:10.1007/978-981-13-0514-6_59

[11] Dronen, N., Foltz, P. W., & Habermehl, K. (2014). Effective sampling for large-scale
automated writing evaluation systems. arXiv preprint arXiv:1412.5659. Fedorov, V. V.
(1972). Theory of optimal experiments. New York: Academic Press.

[12] Hearst, M. (2000). The debate on automated essay grading. IEEE Intelligent Systems and
Their Applications,15(5), 22-37. doi:10.1109/5254.889104

[13] Burstein, J., Kukich, K., Wolfe, S., Lu, C. and Chodorow, M. (1998) 'Enriching
Automated Essay Scoring Using Discourse Marking', in E. Hovy (ed.) Proceedings of
36th Annual Meeting of the Association for Computational Linguistics and 17th
International Conference on Computational Linguistics (ACL-COLING'98) Workshop

on Discourse Relations and Discourse Markers, pp. 15–21, Montréal, Canada.

[14] Crossley, S. A., Greenfield, J., & McNamara, D. S. (2008). Assessing text readability
using cognitively based indices. TESOL Quarterly, 42, 475–493.

[15] Dascalu, M., Dessus, P., Bianco, M., Trausan-Matu, S., & Nardy, A. (2014). Mining
texts, learner productions and strategies with ReaderBench. In Educational Data
Mining (pp. 345-377). Springer, Cham.

[16] Fei Dong, Yue Zhang, and Jie Yang. 2017. Attention-based Recurrent Convolutional
Neural Network for Automatic Essay Scoring. In Proceedings of CONLL. pages 153–
162.

[17] Woods, B., Adamson, D., Miel, S., & Mayfield, E. (2017). Formative Essay Feedback
Using Predictive Scoring Models. Paper presented at the Proceedings of the 23rd ACM
SIGKDD International Conference on Knowledge Discovery and Data Mini

[18] Madnani, N., Burstein, J., Elliot, N., Klebanov, B. B., Napolitano, D., Andreyev, S., &
Schwartz, M. (2018). Writing Mentor: Self-Regulated Writing Feedback for Struggling
Writers. In Proceedings of the 27th International Conference on Computational
Linguistics: System Demonstrations (pp. 113- 117).

[19] Page, E.B. (1994). "New Computer Grading of Student Prose, Using Modern Concepts and Software", Journal of Experimental Education, 62(2), 127-142.