Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 216 (2023) 531–538

Procedia Computer Science

www.elsevier.com/locate/procedia

7th International Conference on Computer Science and Computational Intelligence 2022

Automatic essay exam scoring system : a systematic literature review

Meilia Nur Indah Susanti [a], Arief Ramadhan [a,*], Harco Leslie Hendric Spit Warnars[a]

[a] *Computer Science Department, BINUS Graduate Program – Doctor of Computer Science , Bina Nusantara University, Jakarta, 11480, Indonesia*

Abstract

Currently, Indonesia and the whole world are being hit by the Covid-19 pandemic which has an impact on various fields of life. It affects all sectors, including the education sector. The government through the Ministry of Education and Culture makes a policy in education in terms of the learning process. Teaching and learning activities that were initially carried out face to face become distance learning which was carried out at home. In this study, a systematic literature review is conducted on automatic assessment of essay answers. Various previous studies discuss the essay answer scoring system that has been developed using various methods. We synthesize the results to enrich our understanding of the automated essay exam scoring system. The expected result of this research is that it can contribute to further research related to the automated essay exam scoring system, especially in terms of considering methods and dataset forms.

Keywords: scoring system; essay; automatic; systematic review; assessment

1. Introduction

In 2020, precisely in March, Indonesia (including all parts of the world) experienced a national disaster and was designated as an extraordinary event due to the Covid 19 virus outbreak. The COVID-19 pandemic has greatly

* Corresponding author.
E-mail address: arief.ramadhan@binus.ac.id

10.1016/j.procs.2022.12.166

impacted various fields of life, one of the parts affected is in the field of education [1][2][3][4] Education is a human resource that can bring changes to development and changes to the country. The government, in this case the Ministry of Education and Culture, issued Circular Letter Number 4 of 2020 concerning the Implementation of Education in the Coronavirus Disease (Covid-19) Emergency [5][6]. The circular contains several matters related to policies in the implementation of learning, including the National Examination, School Examination, Class promotion, and the Learning system from home [5].

The policy of the teaching and learning process carried out at home is a new policy that was taken due to the impact of the covid 19 pandemic which caused many changes, education that originally could be done offline (face to face) must be done online (Distance Learning) [7][8][9].

With that instruction, many universities swiftly carry out the instruction that has been set by changing the way of learning. Changes in the learning process also affect the evaluation system process. The evaluation system is one way to measure students' thinking skills, especially in terms of answering questions. To measure this ability, it is necessary to hold an assessment process for the questions that are filled by students. Currently, if learning is carried out face-to-face, the assessment will be carried out directly, but for online learning, the assessment process will be carried out virtually [10][11][12].

In the learning process, assessment is important, where it is usually divided into 3 (three) main parts, namely skills, knowledge, and attitudes. The assessment itself is obtained from the process of scoring a test. In general, the question models given by the teachers are various, for example, questions in the form of multiple choice or essays[13][14]. In automatic essays, the assessment process usually involves several factors to determine the quality of writing which includes content and grammar [15][16].

Computers are a very important tool in the learning process in terms of scoring. Accurate scores will be obtained if the evaluation model is multiple choice, but if the evaluation model is in the form of an essay, the assessment will be complicated due to the wide variety of answers given by students. This impact that resulted in the automatic assessment and correction of essay answers has become the subject of research studies [17][18]. The fact is that more and more students rely heavily on computers to complete and submit their school assignments [19][20][21].

2. Research Methodology

In the process of reviewing this journal, the researcher adopts the Systematic Literature Review (SLR) method described in [22] and [23]. This study seeks to see that the automatic assessment of essay exams is one of the most important things in the learning process. To be able to find out what discussions were discussed by previous research, the we made several steps in the design process and research strategy considering the purpose of this study was to find out the methods used in automatic essay exam assessments.

Following Kitchenham's method [22], we use three stages in the SLR process, ie. planning, conducting a review and reporting the results illustrated in the Figure 1 below:

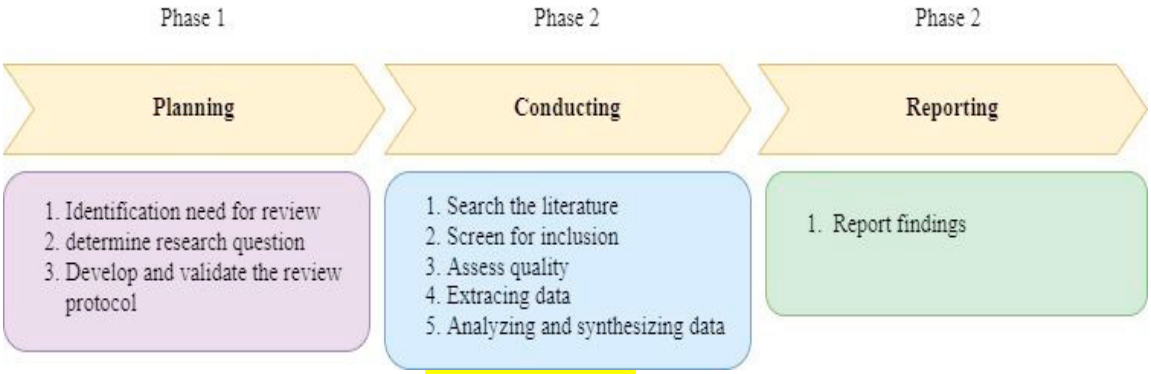

Fig. 1. The SLR process

2.1. *Formulate the problem*

There are three research questions that will be answered through this systematic literature review, ie:

1. R1 : What is the research theme discussed?
2. R2 : What algorithm is used to automatically correct essay exams?
3. R3 : What datasets are used?

2.2. *Develop and validate the review protocol*

We made a series of simple keywords that raised the theme of the Automatic Essay Exam Assessment System using several algorithms to complete essay exam corrections automatically including:

1. Automatic scoring system AND “literature review”
2. Automatic scoring system AND “essay questions”
3. Automatic scoring system AND “algorithm”
4. Automatic scoring system AND “online”
5. Automatic scoring system AND “correction”
6. Automatic scoring system AND “literature survey”

2.3. *Seacrh the literature*

The process of searching for titles, keywords, and abstracts were done by taking several journal papers from Google Scholar. We use Google Scholar because we want to maximize the chance that the paper can be accessed without any burden of license.

In the process of searching for journal papers on Google Scholar that discusses automatic scoring system. The researcher used the keywords: "Automatic scoring system AND "literature review" OR "Automatic scoring system AND "essay questions" OR "Automatic scoring system AND "algorithm" OR "Automatic scoring system AND “online” OR “Automatic scoring system AND “correction” OR “Automatic scoring system AND “literature survey”. From this step we managed to find 33 journal papers. The distribution of the paper found can be seen in Fig. 2.

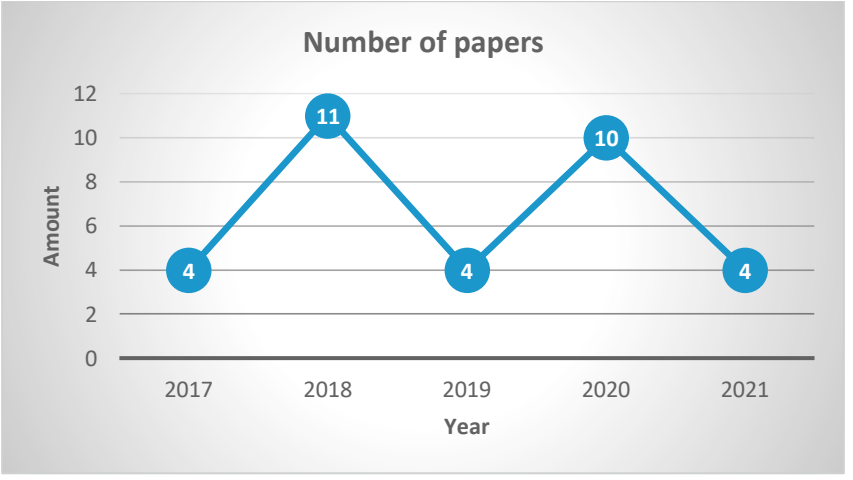

Fig. 2. Related papers found in 2017 – 2021

2.4. *Screen for inclusion*

The inclusion criteria that we used are journal papers in English and Indonesian that have been cited and have a time span between 2017-2021 which discusses automatic correction of essay answers in the field of education. We also

focused on papers that use handwriting as its dataset. Research using a graphic basis is one of the criteria in selecting the obtained papers. Of 33 journal papers obtained, there were only 10 journal papers are used by researchers because the papers fit to our inclusion criteria.

2.5. Assess quality

We assess all reviewed papers based on some suggestions given in [22]. We also consider the research question in our SLR, so the Quality Assessment checklits that we use are:

- Are the research questions or research objectives of the paper clear?
- Does it convey detailed methods?
- Is information about the dataset provided, so that it can be traced?
- Do the conclusions meet the research questions or research objectives?

Based on the results of the Quality Assessment, we found that all papers could meet all of the criteria we provided. Thus, all of the remaining 10 papers are all involved in the next stage.

2.6. Extracting data

At this stage, we conducted data extraction on 10 papers that were found to meet the inclusion criteria. The data that we extracted from each paper are the title of the paper, the name of the author of the paper, the year of publication, the theme raised, the method, and the dataset used. The results of the data extraction of each paper then are used to analyze and find out the processes used in the automatic correction of written exams.

2.7. Analyzing and synthesizing data

We analyzed and synthesized the data that has been extracted from each paper in terms of the method used, and the results obtained from each research. The main goal of this review is to help readers to find out what algorithms are used to automatically correct essay answers in the field of education. We also explore the datasets used and the main themes explored.

2.8. Report findings

In this step, we report the analysis taken by each paper based on the dataset taken, the method used, and a descriptive explanation of the process described in auto-correcting essay exams. The structure of our report is structured in such a way that it can answer the research questions that were asked previously. Our report finding is delivered in the results and discussion.

3. Result and discussion

3.1. Research theme

In [24], the assessment process uses the OpenNLP Framework, the process is carried out by comparing what is entered into the system, it is obtained that essay questions containing symbols and mathematical formulas are the biggest questions in the correction process that can be used for the development of further research.

The results in [25] are to compare their method (Multi-Dimensional Long Short Term Memory and convolution layers) with the results of the use of the Optical Handwriting Recognition (OHR) system with an assessment using Automatic Essay Scoring (AES) both resulted in a Quadratic Weighted Kappa (QWK) score of 0.88, which means that although the current OHR system still has shortcomings, overall it can work well. Whereas the methodology used in [26] is by analyzing starting from text, sub-domains including syntax, semantics, and sentimental are used to improve prediction accuracy, Measurement of predictive suitability produces a QWK value of 0.793.

The research of [27] used Case Folding, Tokenization, Punctuation Removal, Stopword Removal, and Stemming processes. To measure the effectiveness of the pre-processing used, correlation calculations and Mean Absolute Error are used, and there is no significant difference. Handwriting recognition is done in [28] by segmenting words from a

written text sheet scanned into digital form using image processing and the results of the assessment using QWK obtained a score of 0.53 the handwriting reading model works very well because QWK is close to a score of 0.65. On the other hand, the experimental results by [29] are encouraging; the classification range is 85.67 – 91.90% with 100% accuracy, Which means it is more efficient for using Hidden Markov Models (HMM) in Short Answer Questions Systems (SAAS) because the results are promising although, only a small number of samples are required, however, 100% accuracy rate is achieved.

Based on [30], handwriting recognition has not yet reached a state that can directly help scalability automatic evaluation. In order to improve the efficiency and quality of the assessment, a tool for computerizing the assessment of writing is used hand. In other research, handwriting recognition by [31] is %91 correct when given the contents of the answer box.

QWK metrics are used by [32] to train the essay scoring system directly. Long Short-Term Memory (LSTM) is used to assess essay coding and the softmax layer is used to assess essays. On the other hand, handwritten characters scanned images entered in [33] through machine learning classification, namely Convolutional Neural Network (CNN). That proposed system displays the final grade which is classified as 92.86%.

3.2. Method used

The methods used from 10 papers for auto-correction of essay exams in the research reviewed varied in their use. Based on Table 1, papers that use machine learning algorithms are 7 papers including Convolutional Neural Network (CNN), classification, regression, nave Bayes, and Multi-Dimensional Long Short-Term Memory (MDLSTM), the process of changing handwriting into images is discussed in 2 papers reviewed using the Horizontal method. Projection Profile (HPP) and Hough Lines (HL), statistical models can also be used in the process of auto-correcting essay exams discussed in 1 paper using the Discrete Hidden Markov Models (HMMs) method.

Table 1. Method and dataset used

No	Reference Information	Method/tools	Dataset
1	**Manar Joundy Hazar et. al** [24]	Intelligent Tutoring Systems (ITSs)	Dataset from al qadisyiah university students in 2017 collected from web programming courses
2	**Annapurna Sharma and Dinesh Babu Jayagopi** [25]	Multi-Dimensional Long Short Term Memory (MDLSTM) and convolution layers	Dataset of 113 handwritten essays based on the Hewlett Foundation AES dataset
3	**Harneet Kaur Janda et. al** [26]	Rules-based grammar merging and performing initial coherence level checks with grape base	The dataset consists of grades 7 - 10 who have written essays with 150 -550 words from different data which are argumentative, responsive, narrative, persuasive and expository
4	**Uswatun Hasanah, et. al** [27]	Cosine Similarity Method	Using student answers in Indonesian where each answer has been determined by the teacher from 32 students who have answered
5	**Christian Gold and Torsten Zesch** [28]	Recognition model with assessment model using QWK (Quadrated Weighted Kappa)	Students' answers in handwritten form are in a line drawn by the author himself as much as 1300 - 1800 training data and around 250-600 for testing data
6	**Hemmaphan Suwanwiwata et. al** [29]	Discrete Hidden Markov Models (HMMs) and Automated Assessment systems Short Answer Questions (SAAS)	The dataset consists of 3000 - 3400 handwritten samples written in the form of correct answers and incorrect answers
7	**Vijay Rowtula et. al** [30]	Horizontal Projection Profile (HPP) and Hough Lines (HL)	The dataset is obtained from handwriting that has been transferred to image form
8	**Amirali Darvishzadeh et. al** [31]	Livescribe digital pens and draw a box around the final answer, preprocessing	Student answers in handwriting written using a Livescribe digital pen and the writing is in the box
9	**Yucheng Wang et. al** [32]	Classification, regression and ranking, naive bayes	Using the Automated Student Assessent Prize (ASAP) dataset parsed with NLTK
10	**Eman Shaikh et. al** [33]	Convolutional Neural Network (CNN)	Using student data from Mohammad Bin Fahd University as many as 250 images from students' handwriting

3.3. Dataset used

Based on Table 1, the dataset used by the researchers from 10 papers was taken from data from school students in grades 7 to 10 and student data. From a review of several papers, there are 8 papers [24] [25] [26] [27] [28] [29] [31] [33] in the form of handwriting that will be checked automatically, 1 paper [30] using handwritten first converted into an image form which will then be processed by automatic correction and the next 1 paper [32] using the Automated Student Assessment Prize (ASAP) dataset described in Natural Language Toolkit (NLTK).

Theoretically, some studies that examine the automatic assessment of essay answers using a computer technology approach to assess essay questions have been able to assess essay answers either using handwriting, image format, or with answers typed on a computer directly using Automatic Essay Scoring (AES) technology. So that it can be seen the process of human and computer relationships where there is a process of design, evaluation and until implementation is carried out where the assessment process that is carried out automatically can use machine learning models so that this research can be redeveloped using other models of machine learning by developing this AES system to produce many forms and structures of questions.

4. Conclusion

The implementation of an automatic essay correction system in the field of education has been successfully developed with various methods used so that the process of correcting students' or participants' essays is faster and more efficient in terms of time. Of the ten journal papers reviewed, it turns out that many methods or algorithms are used to carry out the correction process to measure the similarity value of students' answers to the answer key so that it can provide scores automatically based on students' essay answers. The data set used from the ten journal papers uses student data which is processed for correction using the algorithms used by the researchers.

Acknowledgements

This work is supported by the Research and Technology Transfer Office, Bina Nusantara University as a part of Bina Nusantara University's International Research Grant contract number: No.017/VR.RTT/III/2021 contract date: 22 March 2021.

References

[1] D. E. Novianti, "Kurikulum dan Pembelajaran di Masa Pandemi Covid 19 Apa dan Bagaimana?," *Pros. Nas. Pendidik. LPPM IKIP PGRI Bojonegoro*, vol. 1, no. 1, pp. 70–75, 2020.

[2] E. F. Kufi, T. Negassa, R. Melaku, and R. Mergo, "Impact of corona pandemic on educational undertakings and possible breakthrough mechanisms," *BizEcons Q.*, vol. 11, no. June, pp. 3–14, 2020, [Online]. Available: https://ideas.repec.org/a/ris/buecqu/0022.html.

[3] A. A. Anjorin, "The coronavirus disease 2019 (COVID-19) pandemic: A review and an update on cases in Africa," *Asian Pac. J. Trop. Med.*, vol. 13, no. 5, pp. 199–203, 2020, doi: 10.4103/1995-7645.281612.

[4] S. Tadesse and W. Muluye, "The Impact of COVID-19 Pandemic on Education System in Developing Countries: A Review," *Open J. Soc. Sci.*, vol. 08, no. 10, pp. 159–170, 2020, doi: 10.4236/jss.2020.810011.

[5] "Kementerian Pendidikan dan Kebudayaan » Republik Indonesia." https://www.kemdikbud.go.id/main/blog/2020/03/se-mendikbud-pelaksanaan-kebijakan-pendidikan-dalam-masa-darurat-penyebaran-covid19 (accessed Jun. 17, 2021).

[6] Rasmitadila *et al.*, "The perceptions of primary school teachers of online learning during the covid-19 pandemic period: A case study in Indonesia," *J. Ethn. Cult. Stud.*, vol. 7, no. 2, pp. 90–109, 2020, doi: 10.29333/ejecs/388.

[7] L. D. Herliandry, Nurhasanah, M. E. Suban, and K. Heru, "Transformasi Media Pembelajaran Pada Masa Pandemi Covid-19," *J. Teknol. Pendidik.*, vol. 22, no. 1, pp. 65–70, 2020, [Online]. Available: http://journal.unj.ac.id/unj/index.php/jtp.

[8] S. Pokhrel and R. Chhetri, "A Literature Review on Impact of COVID-19 Pandemic on Teaching and Learning," *High. Educ. Futur.*, vol. 8, no. 1, pp. 133–141, 2021, doi: 10.1177/2347631120983481.

[9] P. Putra, F. Y. Liriwati, T. Tahrim, S. Syafrudin, and A. Aslan, "The Students Learning from Home Experiences during Covid-19 School Closures Policy In Indonesia," *J. Iqra' Kaji. Ilmu Pendidik.*, vol. 5, no. 2, pp. 30–42, 2020, doi: 10.25217/ji.v5i2.1019.

[10] D. Ratu, A. Uswatun, and H. Pramudibyanto, "Pendidikan Dalam Masa Pandemi Covid-19," *J. Sinestesia*, vol. 10, no. 1, pp. 41–48, 2020, [Online]. Available: https://sinestesia.pustaka.my.id/journal/article/view/44.

[11] R. E. Baticulon *et al.*, "Barriers to Online Learning in the Time of COVID-19: A National Survey of Medical Students in the Philippines," *Med. Sci. Educ.*, vol. 31, no. 2, pp. 615–626, 2021, doi: 10.1007/s40670-021-01231-z.

[12] T. P. M. F. Kelly, "Inequality In Household Adaptation To Schooling Shocks: Covid-Induced Online Learning Engagement In Real Time," *Angew. Chemie Int. Ed. 6(11), 951–952.*, 2020.

[13] T. M. Tashu and T. Horvath, "Pair-wise: Automatic essay evaluation using Word Mover's distance," *CSEDU 2018 - Proc. 10th Int. Conf. Comput. Support. Educ.*, vol. 1, no. Csedu, pp. 59–66, 2018, doi: 10.5220/0006679200590066.

[14] S. Kumar, S. Chakrabarti, and S. Roy, "Earth mover's distance pooling over siamese LSTMs for Automatic short answer grading," *IJCAI Int. Jt. Conf. Artif. Intell.*, vol. 0, no. August, pp. 2046–2052, 2017, doi: 10.24963/ijcai.2017/284.

[15] B. Riordan, A. Horbach, A. Cahill, T. Zesch, and C. M. Lee, "Investigating neural architectures for short answer scoring," pp. 159–168, 2018, doi: 10.18653/v1/w17-5017.

[16] İ. UYSAL and N. DOĞAN, "Automated Essay Scoring Effect on Test Equating Errors in Mixed-format Test," *Int. J. Assess. Tools Educ.*, vol. 8, no. 2, pp. 222–238, 2021, doi: 10.21449/ijate.815961.

[17] O. Iskrenovic-Momcilovic, "Using Computers in Teaching in Higher Education," *Mediterr. J. Soc. Sci.*, vol. 9, no. 4, pp. 71–78, 2018, doi: 10.2478/mjss-2018-0116.

[18] O. Zawacki-Richter, V. I. Marín, M. Bond, and F. Gouverneur, "Systematic review of research on artificial intelligence applications in higher education – where are the educators?," *Int. J. Educ. Technol. High. Educ.*, vol. 16, no. 1, 2019, doi: 10.1186/s41239-019-0171-0.

[19] S. Zhao, Y. Zhang, X. Xiong, A. Botelho, and N. Heffernan, "A memory-Augmented neural model for automated grading," *L@S 2017 - Proc. 4th ACM Conf. Learn. Scale*, pp. 189–192, 2017, doi: 10.1145/3051457.3053982.

[20] G. Kurdi, J. Leo, B. Parsia, U. Sattler, and S. Al-Emari, "A Systematic Review of Automatic Question Generation for Educational Purposes," *Int. J. Artif. Intell. Educ.*, vol. 30, no. 1, pp. 121–204, 2020, doi: 10.1007/s40593-019-00186-y.

[21] M. I. Zulfa, A. Fadli, and Y. Ramadhani, "Classification model for graduation on time study using data mining techniques with SVM algorithm," *AIP Conf. Proc.*, vol. 2094, no. April, 2019, doi: 10.1063/1.5097475.

[22] B. Kitchenham, "Procedures for Performing Systematic Reviews," *Keele Univ. Tech. Rep. TR/SE-0401 ISSN1353-7776*, pp. 240–243, 2004, doi: 10.1145/3328905.3332505.

[23] Y. Xiao and M. Watson, "Guidance on Conducting a Systematic Literature Review," *J. Plan. Educ. Res.*, vol. 39, no. 1, pp. 93–112, 2019, doi: 10.1177/0739456X17723971.

[24] M. J. Hazar, Z. H. Toman, and S. H. Toman, "Automated Scoring for Essay Questions in E-learning," *J. Phys. Conf. Ser.*, vol. 1294, no. 4, 2019, doi: 10.1088/1742-6596/1294/4/042014.

[25] A. Sharma and D. B. Jayagopi, "Automated grading of handwritten essays," *Proc. Int. Conf. Front. Handwrit. Recognition, ICFHR*, vol. 2018-Augus, pp. 279–284, 2018, doi: 10.1109/ICFHR-2018.2018.00056.

[26] H. K. Janda, A. Pawar, S. Du, and V. Mago, "Syntactic, semantic and sentiment analysis: The joint effect on automated essay evaluation," *IEEE Access*, vol. 7, pp. 108486–108503, 2019, doi: 10.1109/ACCESS.2019.2933354.

[27] U. Hasanah, T. Astuti, R. Wahyudi, Z. Rifai, and R. A. Pambudi, "An experimental study of text preprocessing techniques for automatic short answer grading in Indonesian," *Proc. - 2018 3rd Int. Conf. Inf. Technol. Inf. Syst. Electr. Eng. ICITISEE 2018*, pp. 230–234, 2018, doi: 10.1109/ICITISEE.2018.8720957.

[28] C. Gold and T. Zesch, "Exploring the Impact of Handwriting Recognition on the Automated Scoring of Handwritten Student Answers," *Proc. Int. Conf. Front. Handwrit. Recognition, ICFHR*, vol. 2020-Septe, pp. 252–257, 2020, doi: 10.1109/ICFHR2020.2020.00054.

[29] H. Suwanwiwat, A. Das, M. Ferrer, U. Pal, and M. Blumenstein, "An investigation of discrete Hidden Markov Models on handwritten short answer assessment system," *Icprai*, vol. 2, no. September, p. 18, 2018.

[30] V. Rowtula, V. Bhargavan, M. Kumar, and C. V. Jawahar, "Scaling handwritten student assessments with a document image workflow system," *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work.*, vol. 2018-Janua, pp. 2420–2427, 2018.

[31] A. Darvishzadeh, N. Entezari, and T. Stahovich, "Finding the answer: Techniques for locating students' answers in handwritten problem solutions," *Proc. Int. Conf. Front. Handwrit. Recognition, ICFHR*, vol. 2018-Augus, pp. 587–592, 2018, doi: 10.1109/ICFHR-

2018.2018.00108.

[32] Y. Wang, Z. Wei, Y. Zhou, and X. Huang, "Automatic essay scoring incorporating rating schema via reinforcement learning," *Proc. 2018 Conf. Empir. Methods Nat. Lang. Process. EMNLP 2018*, pp. 791–797, 2020, doi: 10.18653/v1/d18-1090.

[33] E. Shaikh, I. Mohiuddin, A. Manzoor, G. Latif, and N. Mohammad, "Automated grading for handwritten answer sheets using convolutional neural networks," *2019 2nd Int. Conf. New Trends Comput. Sci. ICTCS 2019 - Proc.*, pp. 1–6, 2019, doi: 10.1109/ICTCS.2019.8923092.