Hochschule
Bonn-Rhein-Sieg
University of Applied Sciences

Bonn-Aachen
International Center for
Information Technology

Master Thesis Proposal

Computer Assisted Short Answer Grading with Rubrics using Active Learning

Ganesamanian Kolappan

Supervised by

Prof. Dr. Paul G. Plöger
Prof. Dr. Manfred Kaul
M.Sc Tim Metzler

May 2022

1 Introduction

The examination is a practice of assessing students' knowledge in their learning process [9]. This examination might be of any form namely written, oral, practical, or computer-based (higher usage at recent times) [18]. Irrespective of the type of exam the possible question types are fill in the blanks, multiple-choice, short answer, essay, reading comprehension, or others that include math formula, coding, and matching [5, p. 5]. Comparatively, the short answer has gained more interest [8]. Since it has a unique combination of three criteria [5], 1. students have to recall or think from their knowledge instead of recognizing/choosing, 2. the answer is limited to one or two sentences or max one paragraph, and 3. it is close-ended where the content is preferred instead style of writing. Style of writing is focused on essay grading where the flow of content/keywords and organization of information is favoured [7]. The grading can be either human or computer-based. Computer-based grading of short answers is called Automatic Short Answer Grading (ASAG) which is the focus of this research. Sometimes, ASAG requires a human grader to assist in grading in this situation it is semi-automatic or computer assisted [12].

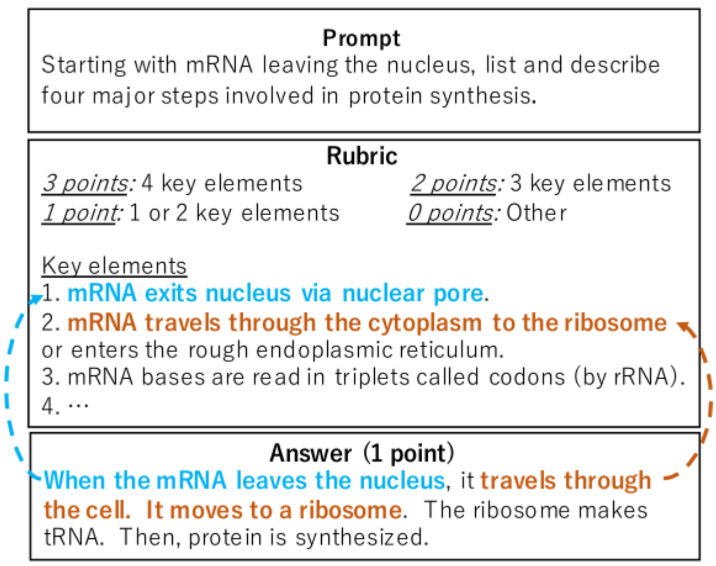

Figure 1: *Example prompt with rubrics from Kaggle's ASAP-SAS dataset provided in [31, p. 1]*

Rubrics are a common assessment technique followed by graders to evaluate students' answers consistently and to provide feedback. Assessment based on rubric highlight the area or topic that the student has to improve [27]. Rubrics state the key elements that need to be present in the answer with their corresponding scores as depicted in Figure 1. This of two types; one is positive rubrics when the key elements are mentioned in the answers their corresponding scores add up to a total score for that particular answer [31, p. 1]. Figure 1 is an example of positive rubrics, since the answer contains two key elements as per the rubrics it scores one point. Another is negative rubrics when the key element is missing their respective scores add up and subtracted from the total score for that particular answer. The key element that is not presented can be provided as feedback (formative assessment) for the students to improve.

Active Learning (AL) is a wrapper that can be placed above any model [29]. AL allows the model to query to a human grader/oracle/annotator to label the data during training [17]. This approach helps in training the model with few labeled data along with a pool of unlabeled data. The unlabeled data are labeled, with the knowledge gained from labeled data. If the model could not label the data it queries to the human grader/oracle/annotator. Thus, active learning helps annotate a large amount of data inexpensively. This process is called semi-supervised learning.

1.1 Motivation

ASAG has been an active research area since 1995 [5]. ASAG has been developed across various domain namely citizenship exams, foreign language learning, classroom exam, entrance exam and general tests comprising different task which includes short answer, essay questions, and reading comprehension [12]. ASAG has several benefits such as

- Grades are available faster, there is no longer a waiting time for students. Additionally, teachers can invest less time in grading where they need to supervise the ASAG [11]
- Grading is consistent whereas human graders may tend to be wrong sometimes due to fatigue, stress, bias or the effects of ordering [5][11][6]

- Grading can be provided for small to a large groups of students [6]
- Grading as well as feedback is available that combines both summative and formative assessments [18][5][6]
- Grading style of the grader can be integrated

The idea of ASAG can be extended to other similar domains which require grading such as an interview or competitive test [32], entrance and certification exams, quiz competitions, or similar. Additionally, ASAG applies to any course ranging from science to computer engineering across different languages including Chinese, English, German, and French [5][11].

AL can be used with any learning methods [20][3],[25]. AL is powerful in working with data of fewer annotations [17]. ASAG with AL wrapper retains the human grader/oracle in the loop for supervising the grades, in parallel, minimizing the effort of graders and maximizing the effectiveness of grading.

1.2 Problem Statement

Presently, there is a need for ASAG for consistent assessment as new questions and different responses are generated regularly [5]. Remarkably, the current ASAG systems are based on a supervised learning method that requires labeled data [12]. Additionally, these model grades are based on the reference answers provided by the grader or automatically selected by the model using clustering [5][18]. These approaches induce difficulties such as:

- The labeled data are annotated manually which is expensive and time-consuming. Sometimes, these data are required in large amounts if the model is deep learning-based supervised learning [6][31]
- Deep supervised learning requires a large amount of data as well as not fast enough to grade as it requires more computational time.
- Having reference answers for each question requires one or two human graders/experts' authorization which is not cost-effective
- No generalization of all ways to answer correctly to a particular question [18]

- Most of the ASAG does not provide feedback
- Partial grades are not in consideration
- Sometimes the supervision of human grader is ignored or ASAG is treated as a replacement of human grader [5]

Current approaches neglect the importance of rubrics, which is significant in a real-world situation for evaluating students' answer[31]. Hence, having rubrics using AL could address the above-mentioned difficulties induced by the present model for ASAG. Research by Marvaniya et al. [18], Wang et al. [31], and Hasanah et al. [11] has proven that including rubrics in ASAG has improved the performance. The Research Questions (RQ) to be answered in this research work are as follows:

RQ1 What are the available methods for ASAG?

RQ2 Does the rubrics aid in providing proper and helpful feedback with grades?

RQ3 Is the model able to generalise?

RQ4 Which models are suitable for AL to grade fast with effectiveness?

2 Related Work

The literature survey is of three parts, namely 1. Literature on ASAG, 2. Literature on ASAG using rubrics, and 3. Literature on AL for ASAG.

Burrows et al. [5] presents a wide range of methods for ASAG explored so far across the years from 1995 to 2015. The research includes some of the notable methods namely, latent semantic analysis [19] one of the familiar methods for ASAG that matches the key terms between the students' answers and provided reference answers, string-based or edit distance based similarity [23] search the resemblance in character or term level, word embedding or word semantic network-based similarity that is WordNet [28][22]. Latterly, deep learning-based similarity representation at the feature level is proved to be effective for ASAG [24][15]. Most of the existing ASAG systems exhibit concern in better representing the similarities between reference answers and student answers.

However, the usage of rubrics instead of reference answers has gained limited attention. Sakaguchi et al. [26] have computed similarities between each key element in rubrics with students' answers which ignores the meaning of the sentences. Marvaniya et al. [18] have used rubrics to provide feedback to their tutoring systems which have limited capabilities. Hasanah et al. [11] have given a basic usage of rubrics for ASAG in the Indonesian language. Also, the paper has mentioned certain tools namely part-of-speech tagging, wordnet is not available for Indonesian language. Wang et al. [31] induces a rubric component with the existing neural-based ASAG architecture which uses the benefit of having both reference answers and rubrics where the computational time might belong.

AL is been used as wrapper for different models, to mention few, random forest [20], convolutional neural network [3], support vector machines [16], reinforcement learning [25] and deep learning [2][3] for different tasks such as time-series classification [2], anomaly detection [16], image classification and detection [4]. Whereas, in educational domain, AL has been used for Arabic text classification [10], [1], Niraula and Rus [21] uses AL for gap-filling questions where the focus is fixed and length is one or two words. Dronen et al. [7] apply AL for automatic essay scoring, Kishaan et al. [13] had a comparison study for ASAG using AL. Horbach and Palmer [12] claims to be the first to employ AL for ASAG.

3 Methodology

This research work combines the idea of Wang et al. [31] and Horbach and Palmer [12] by having rubrics using AL. This could be the first research work that incorporates rubrics using AL. The methodological approach is presented in Figure 2. The dataset consists of around 200 answers for one question with grades from a statistics and probability course taught in German at the Hochschule Bonn-Rhein-Sieg, University of Applied Sciences. This dataset undergoes preprocessing namely case folding, spelling correction, and tokenization. Later, significant features are extracted from these preprocessed data which will be fed as input to the model

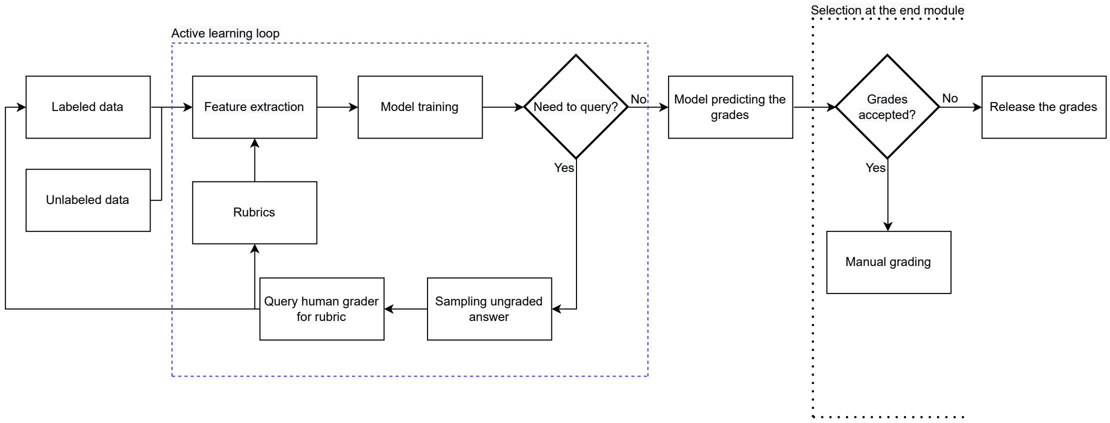

Figure 2: *Methodological approach*

for training. The model under training could be from ensemble methods such as random forest or adaboost wrapped by active learning along with the feature extraction method. Active learning helps to query for rubrics to enhance the performance of the model. Also, the model decides which ungraded answer it should query to a human expert so that it can maximize its performance. When the need to querying is satisfied the model exits the active learning loop and predicts the grades. This grade can be either accepted by the grader or he can choose to manually grade that particular answer in case he is not pleased by the grade provided by ASAG.

3.1 Evaluation

Evaluation of ASAG is done by correlating the grades produced by ASAG and the grades given by the human grader [11]. Success of ASAG is determined by how much similar grades are generated by ASAG to the human grader. This research uses accuracy, F1-score and weighted quadratic kappa as evaluation metrics.

3.1.1 Accuracy

The measure of closeness of the predicted value to the true value is called accuracy. This provides information about the performance of the method [14].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

where, TP is True Positive, TN is True Negative, FN is False Negative, and FP is False Positive.

3.1.2 F1-score

F1-score gives information about the enhancement in the performance of the method. In the classification task, there are two main objectives; one is to minimize the incorrect classification (FP), which maximizes the precision, other is to minimize the incorrect missing(FN) to classify it correctly, which maximizes the recall. These two parameters provide the direction of enhancement [14].

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F_1\text{-}score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{4}$$

3.1.3 Weighted Quadratic Kappa

Kappa or Cohen's Kappa is a measure of agreement between two annotators in a classification task [30]. In this research, the kappa is calculated between the grades generated by the ASAG and grades given by the human grader which can be calculated by

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{5}$$

where, κ is kappa value, p_0 is observed agreement, p_e is expected agreement. Since most of the ASAG are evaluated using quadratic weighted kappa this research uses the same [5][12][31]. Quadratic weighted kappa adds quadratic weights to the agreement value.

4 Project Plan

4.1 Work Packages

The following are the work packages associated with this project, which are to be delivered as a whole package at the end of this project.

Work package	Work package	Task
WP1	**Literature Review**	Study on methods in Automatic Short Answer Grading
		Study on Active learning
WP2	**Dataset Collection and Cleaning**	Dataset collection
		Data preprocessing
WP3	**Feature Extraction**	Study on feature extraction of data
		Extract significant features from data
WP4	**Model Training and Evaluation**	Training model on the data features with active learning wrapper
		Evaluate the model performance
WP5	**Analysis of the Results**	Analysis of the result and fine tuning the parameter to enhance it
WP6	**Final report**	Determine the future works and improvements
		Final report

Figure 3: *Work package*

4.2 Milestones

In order to develop this research project in an organized manner, the project is divided into the following milestones.

M1 Literature review

M2 Dataset finalization

M3 Feature extraction of data

M4 Model implementation with active learning

M5 Experimental results

M6 Report Submission

4.3 Project Schedule

The overall research work target period is provided in Figure 4

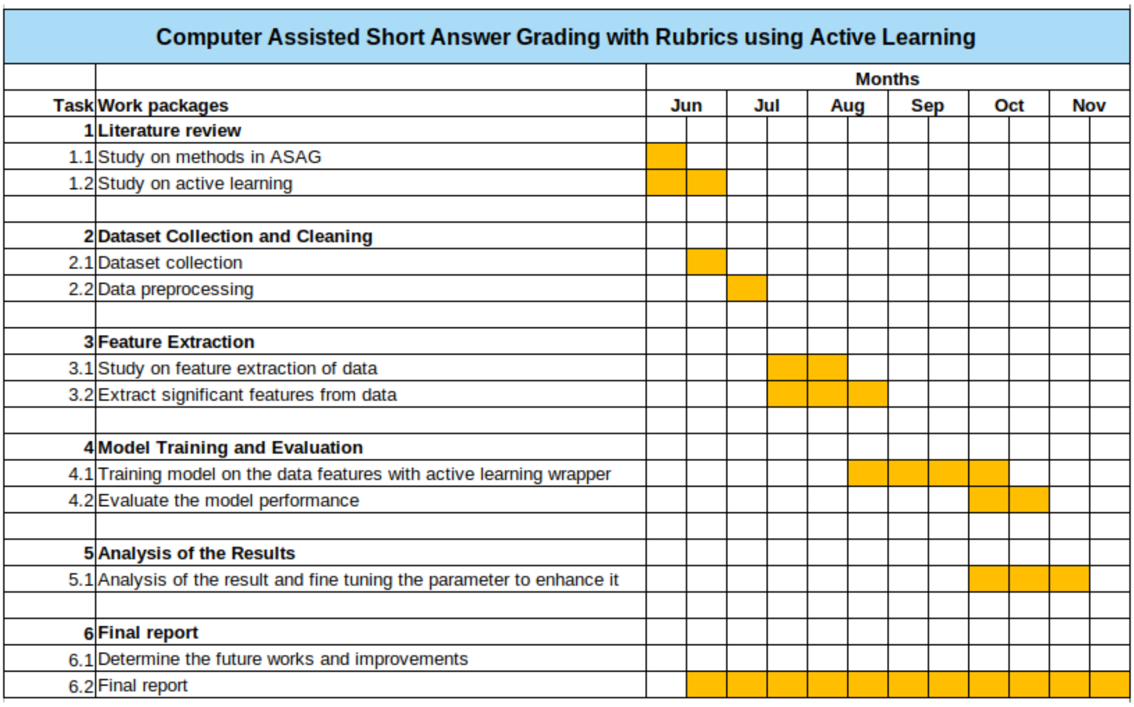

Figure 4: *Research work timeline*

4.4 Deliverables

Minimum Viable

- Literature review
- Analysis of the state of the art
- Dataset collection and analysis
- Model implementation for ASAG with rubrics using AL
- Final report

Expected

- Overview of feature extraction methods

Maximum

- Model comparison for ASAG
- Proposed method, presentable as executable tool or as in interface

References

[1] Abdel-Karim Al-Tamimi, Esraa Bani-Isaa, and Ahmed Al-Alami. Active learning for arabic text classification. In *2021 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)*, pages 123–126, 2021. doi: 10.1109/ICCIKE51210.2021.9410758.

[2] Sergio Barraza, William Lindskog, Davide Badalotti, Oskar Liew, and Arash Toyser. Active learning framework for time-series classification of vibration and industrial process data. *Annual Conference of the PHM Society*, 13, 11 2021. doi: 10.36001/phmconf.2021.v13i1.3059.

[3] Marco Bellini, Georges Pantalos, Peter Kaspar, Lars Knoll, and Luca De-Michielis. An active deep learning method for the detection of defects in power semiconductors. In *2021 32nd Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC)*, pages 1–5, 2021. doi: 10.1109/ASMC51741.2021.9435657.

[4] Lorenzo Bruzzone and Claudio Persello. Active learning for classification of remote sensing images. In *2009 IEEE International Geoscience and Remote Sensing Symposium*, volume 3, pages III–693–III–696, 2009. doi: 10.1109/IGARSS.2009.5417857.

[5] Steven Burrows, Iryna Gurevych, and Benno Stein. The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education*, 25, 03 2015. doi: 10.1007/s40593-014-0026-8.

[6] Aubrey Condor, Max Litster, and Zachary A. Pardos. Automatic short answer grading with sbert on out-of-sample questions. In *EDM*, 2021.

[7] Nicholas Dronen, Peter Foltz, and Kyle Habermehl. Effective sampling for large-scale automated writing evaluation systems. 12 2014. doi: 10.1145/2724660.2724661.

[8] Lucas Galhardi and Jacques Brancher. *Machine Learning Approach for Automatic Short Answer Grading: A Systematic Review*, pages 380–391. 11 2018. ISBN 978-3-030-03927-1. doi: 10.1007/978-3-030-03928-8_31.

[9] Lucas Galhardi, Helen Senefonte, Rodrigo Clemente Thom De Souza, and Jacques Brancher. Exploring distinct features for automatic short answer grading. 10 2018. doi: 10.5753/eniac.2018.4399.

[10] Mohamed Goudjil, Mouloud Koudil, Nacereddine Hammami, Mouldi Bedda, and Meshrif Alruily. Arabic text categorization using svm active learning technique: An overview. In *2013 World Congress on Computer and Information Technology (WCCIT)*, pages 1–2, 2013. doi: 10.1109/WCCIT.2013.6618666.

[11] Uswatun Hasanah, Adhistya Permanasari, Sri Kusumawardani, and Feddy Pribadi. A scoring rubric for automatic short answer grading system. *TELKOMNIKA Indonesian Journal of Electrical Engineering*, 17:763, 04 2019. doi: 10.12928/telkomnika.v17i2.11785.

[12] Andrea Horbach and Alexis Palmer. Investigating active learning for short-answer scoring. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 301–311, San Diego, CA, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-0535. URL `https://aclanthology.org/W16-0535`.

[13] Jeeveswaran Kishaan, Mohandass Muthuraja, Deebul Nair, and Paul G. Plöger. Using active learning for assisted short answer grading. ICML 2020 Workshop on Real World Experiment Design and Active Learning, 2020.

[14] Ganesamanian Kolappan. Feature extraction for motion data. Master's thesis, Hochschule Bonn-Rhein-Sieg, Germany, 2020.

[15] Surya Krishnamurthy, Ekansh Gayakwad, and Nallakaruppan Kailasanathan. Deep learning for short answer scoring. *International Journal of Recent Technology and Engineering*, 7:1712–1715, 03 2019.

[16] Wanchak Lenwari. Genetic algorithms-based gain optimization of a simple learning control for single-phase shunt active filters. In *ICCAS 2010*, pages 2457–2461, 2010. doi: 10.1109/ICCAS.2010.5670263.

[17] Hui Li, Xuejun Liao, and Lawrence Carin. Active learning for semi-supervised multi-task learning. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1637–1640, 2009. doi: 10.1109/ICASSP.2009.4959914.

[18] Smit Marvaniya, Swarnadeep Saha, Tejas I. Dhamecha, Peter Foltz, Renuka Sindhgatta, and Bikram Sengupta. Creating scoring rubric from representative student answers for improved short answer grading. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, page 993–1002, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450360142. doi: 10.1145/3269206.3271755. URL `https://doi.org/10.1145/3269206.3271755`.

[19] Michael Mohler and Rada Mihalcea. Text-to-text semantic similarity for automatic short answer grading. pages 567–575, 01 2009. doi: 10.3115/1609067.1609130.

[20] Hieu T. Nguyen, Joseph Yadegar, Bailey Kong, and Hai Wei. Efficient batch-mode active learning of random forest. In *2012 IEEE Statistical Signal Processing Workshop (SSP)*, pages 596–599, 2012. doi: 10.1109/SSP.2012.6319769.

[21] Nobal Bikram Niraula and Vasile Rus. Judging the quality of automatically generated gap-fill question using active learning. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 196–206, Denver, Colorado, June 2015. Association for Computational Linguistics. doi: 10.3115/v1/W15-0623. URL `https://aclanthology.org/W15-0623`.

[22] Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. Wordnet::similarity - measuring the relatedness of concepts. 04 2004.

[23] Feddy Setio Pribadi, Teguh Bharata Adji, and Adhistya Erna Permanasari. Automated short answer scoring using weighted cosine coefficient. In *2016 IEEE Conference on e-Learning, e-Management and e-Services (IC3e)*, pages 70–74, 2016. doi: 10.1109/IC3e.2016.8009042.

[24] Brian Riordan, Andrea Horbach, Aoife Cahill, Torsten Zesch, and Chong Min Lee. Investigating neural architectures for short answer scoring. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 159–168, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-5017. URL `https://aclanthology.org/W17-5017`.

[25] Hesam Sagha, Saeed Bagheri Shouraki, Hosein Khasteh, and Ali Akbar Kiaei. Reinforcement learning based on active learning method. In *2008 Second International Symposium on Intelligent Information Technology Application*, volume 2, pages 598–602, 2008. doi: 10.1109/IITA.2008.565.

[26] Keisuke Sakaguchi, Michael Heilman, and Nitin Madnani. Effective feature integration for automated short answer scoring. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1049–1054, Denver, Colorado, May–June 2015. Association for Computational Linguistics. doi: 10.3115/v1/N15-1111. URL `https://aclanthology.org/N15-1111`.

[27] Patricia Santos, Xavier Colina, Davinia Hernandez-Leo, Javier Melero, and Josep Blat. Enhancing computer assisted assessment using rubrics in a qti editor. In *2009 Ninth IEEE International Conference on Advanced Learning Technologies*, pages 303–305, 2009. doi: 10.1109/ICALT.2009.92.

[28] Md Arafat Sultan, Cristobal Salazar, and Tamara Sumner. Fast and easy short answer grading with high accuracy. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1070–1075, San Diego, California, June

2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1123. URL `https://aclanthology.org/N16-1123`.

[29] Li-Li Sun and Xi-Zhao Wang. A survey on active learning strategy. In *2010 International Conference on Machine Learning and Cybernetics*, volume 1, pages 161–166, 2010. doi: 10.1109/ICMLC.2010.5581075.

[30] Juan Wang, Yongyi Yang, and Bin Xia. A simplified cohen's kappa for use in binary classification data annotation tasks. *IEEE Access*, 7:164386–164397, 2019. doi: 10.1109/ACCESS.2019.2953104.

[31] Tianqi Wang, Naoya Inoue, Hiroki Ouchi, Tomoya Mizumoto, and Kentaro Inui. Inject rubrics into short answer grading system. pages 175–182, 01 2019. doi: 10.18653/v1/D19-6119.

[32] Muhammad Yusuf and Kemas M Lhaksmana. An automated interview grading system in talent recruitment using svm. In *2020 3rd International Conference on Information and Communications Technology (ICOIACT)*, pages 34–38, 2020. doi: 10.1109/ICOIACT50329.2020.9332109.