Scoring with Confidence? –
Exploring High-confidence Scoring for Saving Manual Grading Effort

Marie Bexte[1] Andrea Horbach[1, 2] Lena Schützler[3]
Oliver Christ[3] Torsten Zesch[1]
[1]CATALPA, FernUniversität in Hagen, Germany
[2]Hildesheim University, Germany
[3]FernUniversität in Hagen, Germany

Abstract

A possible way to save manual grading effort in short answer scoring is to automatically score answers for which the classifier is highly confident. We explore the feasibility of this approach in a high-stakes exam setting, evaluating three different similarity-based scoring methods, where the similarity score is a direct proxy for model confidence. The decision on an appropriate level of confidence should ideally be made before scoring a new prompt. We thus probe to what extent confidence thresholds are consistent across different datasets and prompts. We find that high-confidence thresholds vary on a prompt-to-prompt basis, and that the overall potential of increased performance at a reasonable cost of additional manual effort is limited.

1 Introduction

Whenever a (semi-)automatic process is used to assist humans in scoring free-text answers, there is a trade-off between the human workload required and the resulting scoring accuracy. Without any human input, the accuracy of the automated rating is usually quite low (Egaña et al., 2023), however, already little human input might go a long way in improving the automation quality. Suen et al. (2023) score answers in a setting that uses reference answers and operationalize the confidence of the model as the similarity to the closest reference answer. This concept is visualized in Figure 1. They find that setting a threshold on model confidence, deferring to manual evaluation what falls short of it, leads to reasonable manual effort and high scoring accuracy.

We test the applicability of this method in a high-stakes classroom setting, where items are usually not re-used. This sharply limits the amount of manual scoring effort that can be spent before automation becomes uneconomical. We thus use a small volume of reference answers and examine to what extent a sensible pre-set confidence threshold can be established. As we cannot make the high-stakes student answers publicly available, we additionally replicate our results on four widely used datasets.

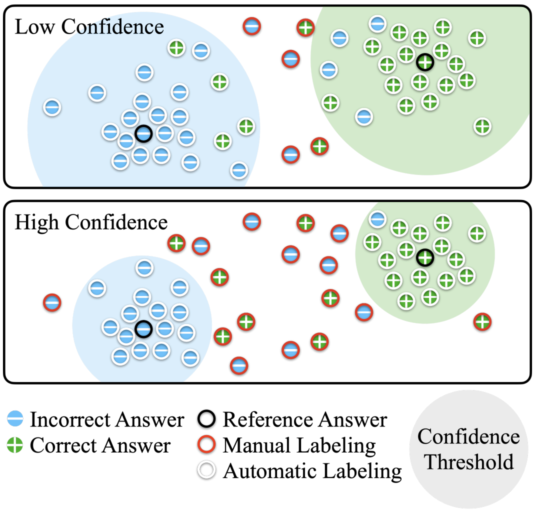

Figure 1: Confidence-based scoring

Our study makes the important step of linking state-of-the-art natural language processing for rating free-text items with the practical questions of start-up costs for building the models.

2 Related Work

The idea to automatically score only parts of all answers or to defer answers with a particularly low confidence of the algorithm to human scoring has been explored before (Funayama et al., 2020, 2022). The approach that is closest to ours is Suen et al. (2023), where answers to medical exam questions are scored using a similarity-based scoring method (Bexte et al., 2022, 2023) and the confidence of the classifier is operationalized through the similarity to the closest reference answer. This method could be taken further to iteratively improve a classifier through those human-labeled low-confidence answers, i.e. using Active Learning (Settles, 2009), as in the scoring domain done by Horbach and Palmer

			Answer averages across prompts			
Dataset	# Prompts	# Labels	#	% Unique	Length	Language
UniversityExams	7	2	544	34	18	German
ASAP	10	3 or 4	2,227	100	239	English
Beetle	56	2	93	100	49	English
SEB	140	2	42	97	64	English
Powergrading	10	2	678	35	25	English

Table 1: Answer and label statistics of the datasets used in our experiments.

(2016) and Kishaan et al. (2020). Such a procedure does however have the disadvantage that human annotators have to annotate small batches of answers over a longer period of time.

Other studies rely on the idea that similar answers should receive the same score. Such a grouping of answers could be reached through surface-level normalization (cf. Zehner et al. (2016)), which reduces orthographic variance, or unsupervised clustering methods operating on the surface level (Horbach et al., 2014; Zesch et al., 2015; Horbach and Pinkal, 2018; Weegar and Idestam-Almquist, 2023), on the semantic level using, e.g. LSA approaches (Zehner et al., 2016; Andersen et al., 2023), or a combination of the two (Basu et al., 2013).

3 Data

We conduct experiments on five datasets (see Table 1). Our high-stakes exam dataset consists of German answers collected from university students as part of their final exam in a statistics class. We refer to this dataset as UniversityExams. It contains 7 prompts that each require a short answer. An exemplary question (translated from German) is *Name the method that is used to estimate the required sample size before an experiment*, where a satisfactory answer would be *a-priori power analysis*. Answers are labeled on a binary scale as either correct or incorrect. Due to the sensitiveness of this data, we can unfortunately not publish it.

We thus also run experiments on four existing, publicly available English datasets, that we use to put results on the exam data into context: The **ASAP**[1] dataset consists of answers to ten prompts from the domains of Biology, Science, and English Language Arts. **Powergrading** (Basu et al., 2013) has answers to ten United States Citizenship Exam questions that were collected from Amazon Mechanical Turk. The Student Response Analysis (SRA) dataset (Dzikovska et al., 2013) is split into two subsets: **Beetle** and **SciEntsBank (SEB)**. Beetle has answers to 56 questions about electricity and electronics, while SciEntsBank contains answers to 150[2] prompts that are from a mix of 15 different science domains. We use the two-way labeled version of the SRA dataset, where answers are classified as correct or incorrect.

4 Experimental Setup

Data Split We split the answers to each prompt into reference and test answers. Our reference answers aim to simulate a teacher manually providing exemplary answers for the different outcome labels. In practice, this would mean a rather small volume of unique examples per label. For each prompt, we thus randomly sample 5 answers per label as references, ensuring that there are no duplicates in this sample. Whenever a similarity metric is fine-tuned on the reference answers, we split them into four answers per label to train and one answer per label to validate.

Classifiers We compare three methods of similarity-based classification that differ with respect to the employed similarity metric. All use a set of reference answers to label the test answers: Based on the respective similarity metric, we predict the label of the most similar reference answer. We compare the following metrics: (i) Edit distance[3] and two variants of cosine similarity based on (ii) pretrained or (iii) fine-tuned SBERT embeddings (Reimers and Gurevych, 2019).[4] For the English datasets, we use the *all-MiniLM-L6-v2* base model, and for the German data the *paraphrase-multilingual-MiniLM-L12-v2* one, both taken from HuggingFace.

[1] https://www.kaggle.com/c/asap-sas/overview

[2] We combine answers from *training* and *unseen questions*. Since our experiments require at least five answers for each label, we can only use 140 prompts.

[3] Determined using the python *Levenshtein* module: https://github.com/rapidfuzz/Levenshtein

[4] We transform edit distance into a noralized similarity for better comparability by computing *1-edit distance* and scaling by length of the longest answer to the respective prompt.

Dataset	Edit	SBERT		target
		pretrained	finetuned	
UniversityExams	.86	.86	.91	.95
ASAP	.46	.43	.50	.60
Beetle	.65	.65	.68	.80
SEB	.68	.65	.71	.80
Powergrading	.87	.92	.93	1.00

Table 2: Weighted F1 results when all test answers are scored fully automated.

To fine-tune SBERT, we follow the approach by Bexte et al. (2022): We train with pairs of answers that are labeled with a similarity label of 1 if both answers have the same score and 0 otherwise. To form these training examples, we pair each training answer with each other training answer. To validate, we pair each validation answer with each training answer. At inference, each test answer is compared to each training and each validation answer, i.e. all reference answers. We train for 30 epochs with a batch size of 8, using an *OnlineContrastiveLoss* and an *EmbeddingSimilarityEvaluator*.

Evaluation We evaluate using weighted F1, reporting averages across all prompts of a dataset.

5 Experiments

First, we report results of a **fully automatic baseline**. In this approach, all test answers are scored automatically, i.e. assigned the label of the most similar reference answer. We then explore **confidence-based scoring**, only scoring instances where similarity exceeds a given threshold automatically. The remaining answers are referred to a human for manual scoring. The fully automatic baseline can be seen as an extreme case of this threshold-based scoring, where the confidence threshold is set so that all classifier decisions are accepted. We speak of a baseline, as introducing a confidence threshold should discard misclassifications and thus increase scoring performance.

5.1 Fully-automated Baseline

Table 2 shows performance of our three scoring methods on the fully-automated baseline, i.e. when all test answers are labeled automatically. It is apparent that some datasets are easier to score than others, with a rather consistent pattern across scoring methods. Particularly the UniversityExams and Powergrading answers are easier to score, which is in part due to the lower percentage of unique answers in these datasets. Overall, there is a slight advantage of the fine-tuned SBERT over the other methods.

5.2 Confidence-based Scoring

Using a similarity-based approach to score answers brings about the benefit of being able to take the similarity on which the classification hinges as a confidence estimate. Suen et al. (2023) were able to increase performance by deferring answers where the model is not confident enough to manual labeling. This requires a predefined threshold that dictates whether to take the predicted label or seek manual labeling. In a practical setting, there should not be a requirement of having to determine this threshold for each new prompt, as this would require substantial amounts of labeled data for the new prompt, thereby diminishing the advantage of automatic evaluation. To assess whether there is such a threshold that is reasonable to assume for new prompts, we analyze how much well-suited thresholds vary between datasets and prompts.

Data-driven Threshold Selection To decide on a suitable threshold for each prompt, we define a target performance for each dataset. These values are listed in Table 2 (under column 'target') and were chosen to push performance around .10 weighted F1 above the fully-automated baseline. Figure 4 in the Appendix shows that performance of the individual prompts in a dataset varies: For some prompts, the target performance was already reached (or surpassed), while others lie beneath it, at times substantially. For these, we calculate the lowest possible threshold value that reaches the target performance[5]. Weighted F1 is then calculated on all answers for which the model's confidence exceeds this threshold, i.e. calculated only on those answers for which the machine-predicted label is taken. Answers that are deferred to manual labeling are excluded from the performance calculation, as they are by definition assumed to be scored correctly.

Figure 2 (blue bars) shows the determined optimal thresholds, with each bar corresponding to a prompt of the respective dataset. The only dataset where thresholds are somewhat close together is edit distance-based scoring of Powergrading, where they range from .92 to .99. Otherwise, thresholds vary widely, indicating that it is difficult to predefine a threshold to apply to a new prompt. On top

[5] Prompts already at target level have a threshold of 0.

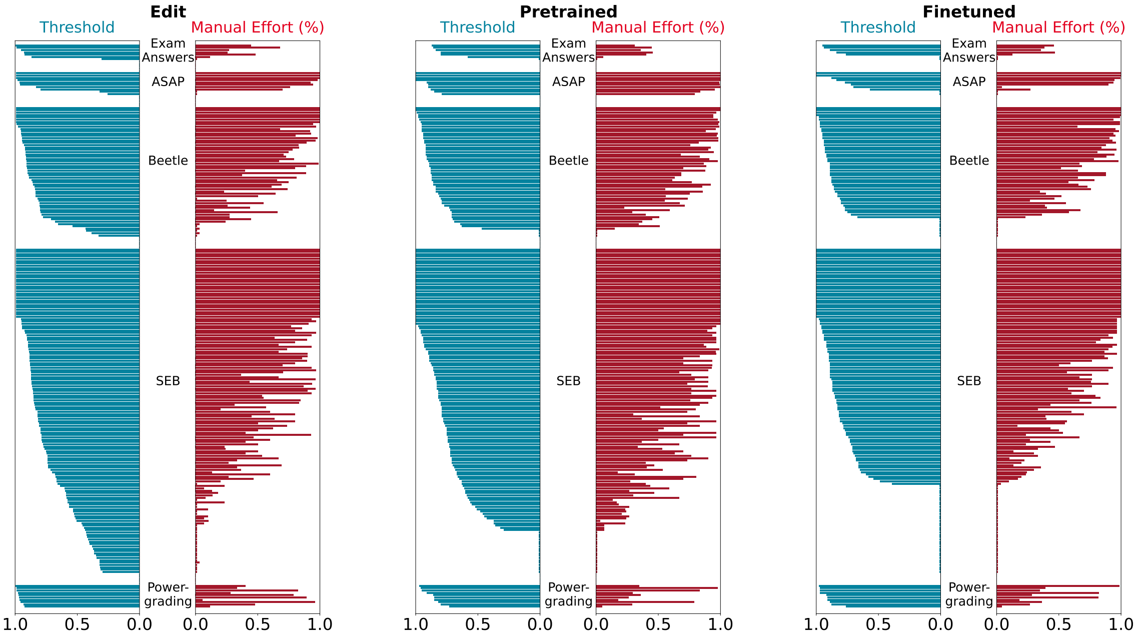

Figure 2: Prompt-wise depiction of thresholds that would have to be set in order to achieve the target performance level (see Table 2). Red bars indicate how much test data falls below the threshold, i.e. has to be scored manually.

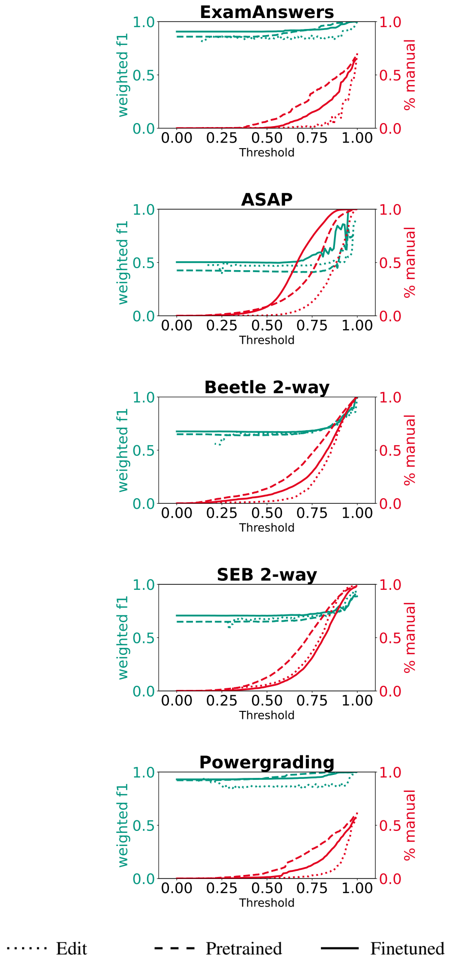

Figure 3: Weighted F1 and amount of answers that requires manual scoring averaged over all prompts of the respective dataset.

of the rather wide span of optimal thresholds, the red bars depict how much of the test data would be deferred to manual labeling. We see that for many of the threshold values, this would make up a substantial amount of answers, often over half of them. Thus, even if there is a threshold found, reaching the target performance level comes at the cost of a large volume of manual annotation effort.

Predefining Threshold Values Instead of a data-driven search for an optimal threshold value, one could also make a top-down decision on a reasonably seeming threshold. Our next analysis inspects how threshold values are related to performance and manual correction effort. Figure 3 shows the relation between threshold value, performance and manual effort averaged over all prompts of a dataset. In general, performance tends to be stable for a rather wide range of thresholds, and only starts to increase when substantial manual effort is required. There is thus no general potential of increasing performance at a reasonable cost of additional manual labeling.

6 Conclusion

While previous work showed that confidence-based scoring can be successful (Suen et al., 2023), we do not find this to hold in our experiments. This may in part be due to the lower volume of reference answers and the higher overall scoring difficulty of some of the datasets we use. On some prompts, there may be thresholds that lead to a desirable tradeoff between manual effort and performance increase, but we did not find a general range of threshold values that would be promising to apply to unseen prompts.

Limitations

Due to the sensitive nature of the exam data, we can unfortunately not publish it. This limits the reproducibility of our results.

When we set thresholds on the similarity, we calculate performance based on only those examples that exceed the confidence threshold. One could also argue to include the answers that are deferred to manual labeling as correctly classified examples. This would increase performance, but it would also mean that a certain volume of answers might be scored with substantially inferior performance, as it would enable for manually labeled answers to even out misclassifications by a model. In practice we want to guarantee a certain level of performance for all students, and hence calculate performance solely on those answers that are classified by a model.

Ethical Considerations

The motivation for this work was to assess the usefulness of automated confidence-based scoring in a high-stakes setting. The performance levels on the SRA and ASAP datasets are however a long way off from being reliable enough for employment in an actual classroom. Even on the better-performing Powergrading and UniversityExams data, the local legal situation is likely to put significant conditions on the use of automated decisions, or even prohibit this entirely.

Achnowledgements

This work was partially conducted at “CATALPA - Center of Advanced Technology for Assisted Learning and Predictive Analytics” of the FernUniversität in Hagen, Germany.

References

Nico Andersen, Fabian Zehner, and Frank Goldhammer. 2023. Semi-automatic coding of open-ended text responses in large-scale assessments. *Journal of Computer Assisted Learning*, 39(3):841–854.

Sumit Basu, Chuck Jacobs, and Lucy Vanderwende. 2013. Powergrading: a clustering approach to amplify human effort for short answer grading. *Transactions of the Association for Computational Linguistics*, 1:391–402.

Marie Bexte, Andrea Horbach, and Torsten Zesch. 2022. Similarity-based content scoring - how to make S-BERT keep up with BERT. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 118–123, Seattle, Washington. Association for Computational Linguistics.

Marie Bexte, Andrea Horbach, and Torsten Zesch. 2023. Similarity-based content scoring-a more classroom-suitable alternative to instance-based scoring? In *Findings of the association for computational linguistics: Acl 2023*, pages 1892–1903.

Myroslava Dzikovska, Rodney Nielsen, Chris Brew, Claudia Leacock, Danilo Giampiccolo, Luisa Bentivogli, Peter Clark, Ido Dagan, and Hoa Trang Dang. 2013. SemEval-2013 Task 7: The Joint Student Response Analysis and 8th Recognizing Textual Entailment Challenge. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 263–274, Atlanta, Georgia, USA. Association for Computational Linguistics.

Aner Egaña, Itziar Aldabe, and Oier Lopez de Lacalle. 2023. Exploration of annotation strategies for automatic short answer grading. In *Artificial Intelligence in Education*, pages 377–388, Cham. Springer Nature Switzerland.

Hiroaki Funayama, Shota Sasaki, Yuichiroh Matsubayashi, Tomoya Mizumoto, Jun Suzuki, Masato Mita, and Kentaro Inui. 2020. Preventing critical scoring errors in short answer scoring with confidence estimation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 237–243.

Hiroaki Funayama, Tasuku Sato, Yuichiroh Matsubayashi, Tomoya Mizumoto, Jun Suzuki, and Kentaro Inui. 2022. Balancing cost and quality: an exploration of human-in-the-loop frameworks for automated short answer scoring. In *International Conference on Artificial Intelligence in Education*, pages 465–476. Springer.

Andrea Horbach and Alexis Palmer. 2016. 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.

Andrea Horbach, Alexis Palmer, and Magdalena Wolska. 2014. Finding a tradeoff between accuracy and rater's workload in grading clustered short answers. In *LREC*, pages 588–595.

Andrea Horbach and Manfred Pinkal. 2018. Semi-supervised clustering for short answer scoring. In

Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).

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

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Burr Settles. 2009. Active learning literature survey.

King Yiu Suen, Victoria Yaneva, Le An Ha, Janet Mee, Yiyun Zhou, and Polina Harik. 2023. ACTA: Short-answer grading in high-stakes medical exams. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 443–447, Toronto, Canada. Association for Computational Linguistics.

Rebecka Weegar and Peter Idestam-Almquist. 2023. Reducing workload in short answer grading using machine learning. *International Journal of Artificial Intelligence in Education*, pages 1–27.

Fabian Zehner, Christine Sälzer, and Frank Goldhammer. 2016. Automatic coding of short text responses via clustering in educational assessment. *Educational and psychological measurement*, 76(2):280–303.

Torsten Zesch, Michael Heilman, and Aoife Cahill. 2015. Reducing annotation efforts in supervised short answer scoring. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 124–132.

A Detailed Results of Fully-automated Baseline

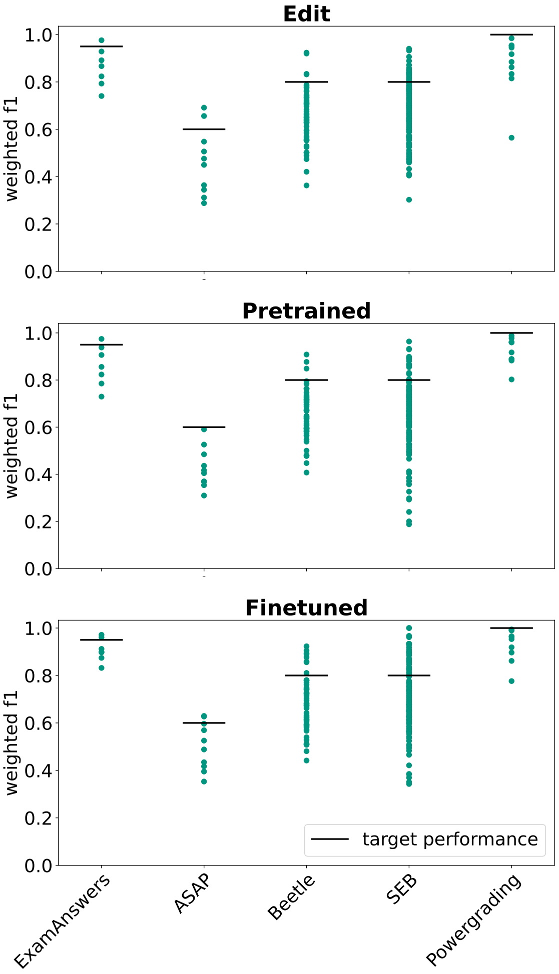

Figure 4: Prompt-wise results for the fully automated baseline and target performance for the respective datasets.