PapagAI: Automated Feedback for Reflective Essays

Veronika Solopova[1 [0000−0003−0183−9433]], Eiad Rostom[1], Fritz Cremer[1], Adrian Gruszczynski[1], Sascha Witte[1], Chengming Zhang[2 [0009−0007−8695−5455]], Fernando Ramos López[1], Lea Plößl[2 [0009−0004−7290−5068]], Florian Hofmann[2], Ralf Romeike[1 [0000−0002−2941−4288]], Michaela Gläser-Zikuda[2 [0000−0002−3071−2995]], Christoph Benzmüller[2,3 [0000−0002−3392−3093]], and Tim Landgraf[1 [0000−0003−4951−5235]]

[1] Freie Universität Berlin, Germany
[2] Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
[3] Otto-Friedrich-Universität Bamberg, Germany

Abstract. Written reflective practice is a regular exercise pre-service teachers perform during their higher education. Usually, their lecturers are expected to provide individual feedback, which can be a challenging task to perform on a regular basis. In this paper, we present the first open-source automated feedback tool based on didactic theory and implemented as a hybrid AI system. We describe the components and discuss the advantages and disadvantages of our system compared to the state-of-art generative large language models. The main objective of our work is to enable better learning outcomes for students and to complement the teaching activities of lecturers.

Keywords: Automated feedback · Dialogue · Hybrid AI · NLP

1 Introduction

Dropout rates as high as 83% among pre-service teachers and associated teacher shortages are challenging the German education system [2,20]. This may be due to learning environments not adequately supporting prospective teachers in their learning process [29]. Written reflective practice may alleviate the problem: By reflecting on what has been learned and what could be done differently in the future, individuals can identify areas for improvement. However, instructors may be overburdened by giving feedback to 200+ students on a weekly basis. With the rise of large language models (LLMs, [30]), automated feedback may provide welcome relief. Students could iteratively improve their reflection based on the assessment of a specialized model and through that, their study performance. Instructors could supervise this process and invest the time saved in improving the curriculum. While current research is seeking solutions to align the responses of LLMs with a given set of rules, it is currently impossible to guarantee an output of a purely learnt model to be correct. Here, we propose "PapagAI", a platform

to write reflections and receive feedback from peers, instructors and a specialized chatbot. PapagAI uses a combination of ML and symbolic components, an approach known as hybrid AI [10]. Our architecture is based on various natural language understanding modules[4], which serve to create a text and user profile, according to which a rule-based reasoner chooses the appropriate instructions.

2 Related work

PapagAI employs a number of models for detecting topics contained in -, and assessing the quality and depth of the reflection, as well as for detecting the sentiment and emotions of the author. While extensive previous work was published on each of these tasks, implementations in German are rare. To our knowledge, there is no previous work that combined all in one application. Automated detection of reflective sentences and components in a didactic context has been described previously [12,18,24,38,36,22]. In [18], e.g., the authors analyse the depth of a reflection on the text level according to a three-level scheme (none, shallow, deep). Document-level prediction, however, can only provide coarse-grained feedback. Liu et al. [23], in contrast, also use three levels for predicting reflective depth for each sentence. In emotion detection, all previous works focus on a small set of 4 to 6 basic emotions. In Jena [16], e.g., the author describes detecting students' emotions in a collaborative learning environment. Batbaatar et al. [1] describes an emotion model achieving an F1 score of 0.95 for the six basic emotions scheme proposed by Ekman [9]. Chiorrini et al. [7] use a pre-trained BERT to detect four basic emotions and their intensity from tweets, achieving an F1 score of 0.91. We did not find published work on the German language, except for Cevher et al. [5], who focused on newspaper headlines. With regard to sentiment polarity, several annotated corpora were developed for German [34,37], mainly containing tweets. Guhr et al. [15] use these corpora to fine-tune a BERT model. Shashkov et el. [33] employ sentiment analysis and topic modelling to relate student sentiment to particular topics in English. Identifying topics in reflective student writing is studied by Chen et al. [6] using the MALLET toolkit [28] and by De Lin et al. [8] with Word2Vec + K-Means clustering. The techniques in these studies are less robust than the current state-of-art, such as ParlBERT-Topic-German [19] and Bertopic [14]. Overall, published work on automated feedback to student reflections is scarce, the closest and most accomplished work being AcaWriter [21] and works by Liu and Shum [23]. They use linguistic techniques to identify sentences that communicate a specific rhetorical function. They also implement a 5-level reflection depth scheme and extract parts of text describing the context, challenge and change. The feedback guides the students to the next level of reflective depth with a limited number of questions. In their user study, 85.7% of students perceived the tool positively. However, the impact on the reflection quality over time was not measured and remains unclear.

[4] All ML models are available in our OSF depository (https://osf.io/ytesn/), while linguistic processing code can be shared upon request.

3 Methods, components and performances

Data collection. Our data comes from the German Reflective Corpus [35]. The dataset contains reflective essays collected via google-forms from computer science and ethics of AI students in German, as well as e-portfolio diaries describing school placements of teacher trainees from Dundee University. For such tasks as reflective level identification and topic modelling, we enlarged it by computer science education students' essays and pedagogy students' reflections[5]. It consists of reflections written by computer science, computer science education, didactics and ethics of AI students in German and English. Data is highly varied, as didactics students write longer and deeper reflections than e.g. their computer science peers.

Emotions detection. Setting out from the Plutchik wheel of basic emotions [31], during the annotation process we realised that many of the basic emotions are never used, while other states are relevant to our data and the educational context (e.g. confidence, motivation). We framed it as a multi-label classification problem at the sentence level. We annotated 6543 sentences with 4 annotators. The final number of labels is 17 emotions, with the 18th label being 'no-emotion'.We calculated the loss using binary cross entropy, where each label is treated as a binary classification problem, the loss is calculated for each label independently, which we sum for the total loss. We achieved the best results with a pre-trained RoBERTa [25] , with a micro F1 of 0.70 and a hamming score of 0.67 across all emotion labels. The model achieved the highest scores for "surprise", "approval" and "interest". With a lenient hamming score, accounting for the model choosing similar emotions (e.g. disappointment instead of disapproval) our model achieves up to 0.73.

Gibbs cycle. [13] illustrates cognitive stages needed for optimal reflective results. It includes 6 phases: *description*, *feelings*, *evaluation*, *analysis*, *conclusion* and *future plans*. We annotated the highest phase present in a sentence and all the phases present. We treated this as a multi-class classification problem and used a pre-trained ELECTRA model. While evaluating, we compared one-hot prediction to the highest phase present and 3 top probability classes with all the phases present. While one-hot matching only managed to score 65% F1 macro, the top 3 predictions achieve up to 98% F1 macro and micro.

Reflective level detection. Under the supervision of Didactics specialists two annotators labelled 600 texts according to Fleck & Fitzpatrick's scheme [11], achieving moderate inter-annotators agreement of 0.68. The coding scheme includes 5 levels: *description*, *reflective description*, *dialogical reflection*, *transformative reflection* and *critical reflection*; With 70% of the data used for the training and 30% for evaluation, we used pre-trained BERT large and complete document embeddings for the English and German, resulting in QWK score of 0.71 in cross-validation.

[5] This still non-published data can be obtained upon request.

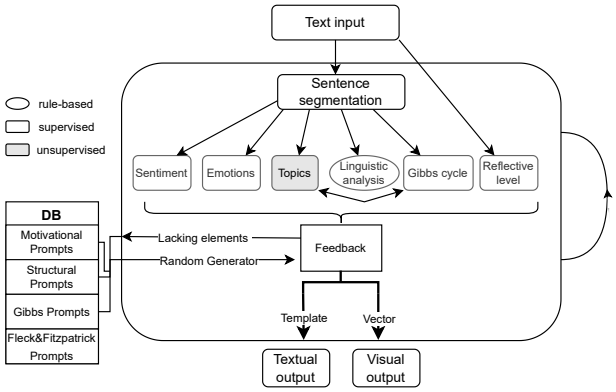

Fig. 1. The diagram of our PapagAI system shows the main productive modules. The legend on the left indicates the nature of the AI modules used.

Topic modelling. We used BERTopic [14] on the sentence level. First, we tokenized and normalize the input sequence to lowercase and filter out numbers, punctuation, and stop-words using nltk library [3]. Then, we extract embeddings with BERT, reduce dimensionalities with UMAP, cluster reduced embeddings with HDBSCAN, create topic representation with tfidf and fine-tune topic representations with the BERT model. Because we have a lot of data of different origins, we created two clusterings, one more specific to the pedagogy topic and one including various educational topics. You can see our clusters in App.

Linguistic scoring. Using spacy[6] we tokenized, and lemmatize the sentences, extracted dependencies parcing and part of speech. Additionally, we used RFTagger[32] for parts of speech and types of verbs. We extract sentence length, adverb for verb ratio, adjective for noun ratio, number of simple and complex sentences, types of subordinated clauses and number of discourse connectors[7] used. This information enables us to determine the reflection length, expressivity and variability of the language, as well as surface coherence and structure.

4 System architecture

In PapagAI (see Fig. 1) the input text of the reflection is received from the AWS server through a WebSocket listener script. To minimize the response time, the models are loaded in the listener script once and then the user request spawn threads with the models already loaded. If the input text is smaller than three sentences and contains forbidden sequences, the processing does not start and the user receives a request to revise their input. Otherwise, the text is segmented

[6] https://spacy.io
[7] We use Connective-Lex list for German: https://doi.org/10.4000/discours.10098.

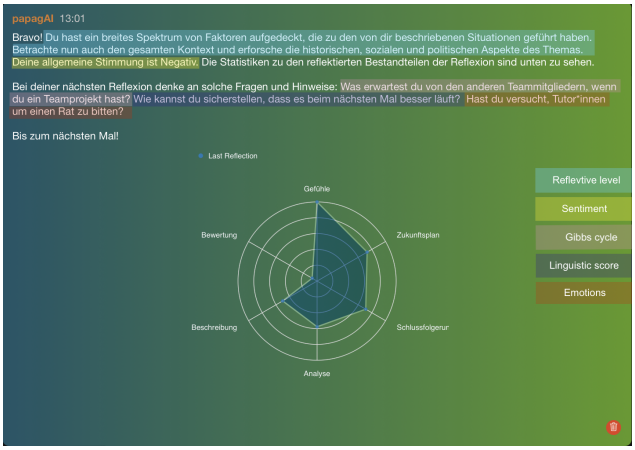

Fig. 2. The radar below the textual feedback illustrates Gibbs cycle completeness. The colour of the highlighted feedback text corresponds to the model responsible for this information.

into sentences and tokens. The language is identified using langid [26] and if the text is not in German, it is translated using Google translator API implementation.[8] The reflective level model receives the whole text, while other models are fed with the segmented sentences. Topic modelling and Gibbs cycle results are mapped, to identify if topics were well reflected upon. If more than three sentences are allocated to the topic and these sentences were identified by the Gibbs cycle model as analysis, we consider the topics well thought through. The extracted features are then passed to the feedback module. Here, the lacking and under-represented elements are identified in linguistic features and the three least present Gibbs cycle stages. If sentiment and emotions are all positive we conclude that no potential challenges and problems are thought through. If the sentiment and emotions are all negative, we want to induce optimism. These features together with the reflective level are mapped to the database of potential prompts and questions, where one of the suitable feedback options is chosen randomly for the sake of variability. Using manually corrected Gpt-3 outputs, for each prompt we created variations so that the feedback does not repeat often even if the same prompts are required.The extracted textual prompts are built together in a rule-based way into the template, prepared for German, Spanish and English. Otherwise, the overall feedback is made in German and then translated into the input language. The textual and a vector of extracted features for visual representation are sent back to the AWS server. The whole processing takes from 15 to 30 seconds based on the length of the text. Sample feedback can be seen in Figure 2.

[8] https://pypi.org/project/deep-translator/

5 Comparison with GPT-3

We compared our emotions detection (fine-tuned RoBERTa) and Gibbs cycle model (fine-tuned Electra) with the prompt-engineered state-of-the-art generative model Davinci [4] on the same task. For the evaluation and comparison, we used a small subset of 262 samples which were not part of the training. We first tried the zero-shot approach, where we described our labels to GPT-3 and gave it our sentence to predict. Then, we tried a one-shot approach, providing GPT-3 with one example sentence for each label. Finally, in the few-shot approach, we provided GPT-3 with three examples per label, which is the maximum number of examples possible due to the input sequence length restriction. Although the task requested GPT-3 to pick multiple labels out of the possible options, the model predicted multiple labels only in 5% of the cases for emotions. For this reason, we used the custom defined "one correct label": the score considers the prediction correct if it contains at least one correct label from the sentence's true labels. The zero-shot approach achieved only 0.28 accuracy in predicting one correct label for emotions. The model predicted the labels "information", "uncertainty", "interest", and "motivated" for the majority of the sentences. With the Gibbs cycle task, it achieved 80% correct predictions. Providing one example per label improved the performance noticeably by 18% (0.46) for emotions, and the model was able to detect emotions like "confidence", "challenged", and "approval" more accurately. It did not influence Gibb's cycle performance. Increasing the number of examples to three resulted in a slight improvement of 3% (0.49) for emotions, and 7% (0.87) for the Gibbs cycle. However, the best-scoring approaches did not offer a comparable performance to our fine-tuned models on these specific tasks with 0.81 on the same custom metric for emotion detection and 0.98 for the Gibbs cycle.

6 Discussion and conclusion

The current PapagAI system has several advantages in comparison to generative LLMs. It ensures transparency of the evaluation and control over the output, which is based exclusively on didactic theory. Although LLMs show huge promise, they are still prone to hallucination [17,27], and, as we have shown in §5, they may under-perform on difficult cognitive tasks in comparison to smaller language models fine-tuned for the task. The fine-tuning of LLMs to didactic books and instructions, which we plan for our future work, still does not guarantee 100% theoretical soundness of the output, which is problematic e.g. in the case of pre-service students with statistically low AI acceptance. At the same time, the newest models, such as GPT-4, are only available through APIs, which raises concerns about data privacy, especially as the data in focus is an intimate reflective diary. Moreover, current open-source models, such as GPT-J and GPT-2, especially for languages other than English do not draw comparable results. Our architecture has, however, obvious drawbacks. On the one hand, our models do not reach 100% accuracy and this can naturally lead to suboptimal feedback.

The processing time for many models, especially for longer texts, can be significantly higher than for a single generative LLM. For now, as we provide one feedback message for one rather long reflection, this is not a big issue, however, if we implement a dialogue form, the time of response would not feel natural. Finally, the variability of output using our approach is much more limited in comparison to generative models. We try to address it by creating many similar versions of instructions rephrased by GPT-3, and corrected manually. On average 7 out of 10 prompts needed some correction. Most of the errors were related to GPT-3 trying to rephrase the given sentence using synonyms that were not didactically appropriate in the given context. Future work, among others, will focus on user studies to understand how we can optimize the feedback, so that the users find it credible and useful, while their reflective skills advance. We also plan a more detailed evaluation based on more user data. We hope that our work will contribute to the optimization of the pre-service teachers' reflective practice and self-guided learning experience.

References

1. Batbaatar, E., Li, M., Ryu, K.H.: Semantic-emotion neural network for emotion recognition from text. IEEE Access **7**, 111866–111878 (2019). https://doi.org/10.1109/ACCESS.2019.2934529
2. Becker, A.: 83 Prozent der Studenten brechen Lehramts-Studium ab. Nordkurier (2021)
3. Bird, S., Klein, E., Loper, E.: Natural language processing with Python: analyzing text with the natural language toolkit. " O'Reilly Media, Inc." (2009)
4. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners (2020)
5. Cevher, D., Zepf, S., Klinger, R.: Towards multimodal emotion recognition in german speech events in cars using transfer learning (2019)
6. Chen, Y., Yu, B., Zhang, X., Yu, Y.: Topic modeling for evaluating students' reflective writing: A case study of pre-service teachers' journals. In: Proceedings of the Sixth International Conference on Learning Analytics & Knowledge. p. 1–5. LAK '16, Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2883851.2883951
7. Chiorrini, A., Diamantini, C., Mircoli, A., Potena, D.: Emotion and sentiment analysis of tweets using bert. In: EDBT/ICDT Workshops (2021)
8. De Lin, O., Gottipati, S., Ling, L.S., Shankararaman, V.: Mining informal & short student self-reflections for detecting challenging topics – a learning outcomes insight dashboard. In: 2021 IEEE Frontiers in Education Conference (FIE). pp. 1–9 (2021). https://doi.org/10.1109/FIE49875.2021.9637181
9. Ekman, P.: Basic emotions. andbook of cognition and emotion **98**, 16 (2023)
10. Elands, P., Huizing, A., Kester, J., Peeters, M.M.M., Oggero, S.: Governing ethical and effective behaviour of intelligent systems: A novel framework for meaningful human control in a military context. Militaire Spectator **188**(6), 302–313 (2019)

11. Fleck, R., Fitzpatrick, G.: Reflecting on reflection: Framing a design landscape. In: Proceedings of the 22nd Conference of the Computer-Human Interaction Special Interest Group of Australia on Computer-Human Interaction. p. 216–223. OZCHI '10, Association for Computing Machinery, New York, NY, USA (2010). https://doi.org/10.1145/1952222.1952269, https://doi.org/10.1145/1952222.1952269
12. Geden, M., Emerson, A., Carpenter, D., Rowe, J.P., Azevedo, R., Lester, J.C.: Predictive student modeling in game-based learning environments with word embedding representations of reflection. Int. J. Artif. Intell. Educ. **31**, 1–23 (2021)
13. Gibbs, G., Unit, G.B.F.E.: Learning by Doing: A Guide to Teaching and Learning Methods. FEU. Oxford Brookes University, Oxford (1988)
14. Grootendorst, M.R.: Bertopic: Neural topic modeling with a class-based tf-idf procedure. ArXiv (2022)
15. Guhr, O., Schumann, A.K., Bahrmann, F., Böhme, H.J.: Training a broad-coverage german sentiment classification model for dialog systems. In: Proceedings of the 12th Language Resources and Evaluation Conference. pp. 1627–1632. European Language Resources Association, Marseille, France (May 2020), https://aclanthology.org/2020.lrec-1.202
16. Jena, R.K.: Sentiment mining in a collaborative learning environment: capitalising on big data. Behaviour & Information Technology **38**(9), 986–1001 (2019). https://doi.org/10.1080/0144929X.2019.1625440
17. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y.J., Madotto, A., Fung, P.: Survey of hallucination in natural language generation. ACM Computing Surveys **55**(12), 1–38 (mar 2023). https://doi.org/10.1145/3571730
18. Jung, Y., Wise, A.F.: How and how well do students reflect?: multi-dimensional automated reflection assessment in health professions education. Proceedings of the Tenth International Conference on Learning Analytics & Knowledge (2020)
19. Klamm, C., Rehbein, I., Ponzetto, S.: Frameast: A framework for second-level agenda setting in parliamentary debates through the lense of comparative agenda topics. ParlaCLARIN III at LREC2022 (2022)
20. Klemm, K., Zorn, D.: Steigende Schülerzahlen im Primarbereich: Lehrkräftemangel deutlich stärker als von der KMK erwartet. Bertelsmann Stiftung (09 2019)
21. Knight, S., Vijay Mogarkar, R., Liu, M., Kitto, K., Sandor, A., Lucas, C., Wight, R., Sutton, N., Ryan, P., Gibson, A., Abel, S., Shibani, A., Buckingham Shum, S.: Acawriter: A learning analytics tool for formative feedback on academic writing. Journal of Writing Research **12**(1), 141–186 (Jun 2020). https://doi.org/10.17239/jowr-2020.12.01.06
22. Kovanović, V., Joksimović, S., Mirriahi, N., Blaine, E., Gašević, D., Siemens, G., Dawson, S.: Understand students' self-reflections through learning analytics. In: Proceedings of the 8th International Conference on Learning Analytics and Knowledge. p. 389–398. LAK '18, Association for Computing Machinery, New York, NY, USA (2018). https://doi.org/10.1145/3170358.3170374
23. Liu, M., Kitto, K., Buckingham Shum, S.: Combining factor analysis with writing analytics for the formative assessment of written reflection. Computers in Human Behavior **120**, 106733 (2021). https://doi.org/https://doi.org/10.1016/j.chb.2021.106733
24. Liu, M., Shum, S.B., Mantzourani, E., Lucas, C.: Evaluating machine learning approaches to classify pharmacy students' reflective statements. In: Isotani, S., Millán, E., Ogan, A., Hastings, P.M., McLaren, B.M., Luckin, R. (eds.) Artificial Intelligence in Education - 20th International Conference, AIED 2019, Chicago, IL,

USA, June 25-29, 2019, Proceedings, Part I. Lecture Notes in Computer Science, vol. 11625, pp. 220–230. Springer (2019). https://doi.org/10.1007/978-3-030-23204-7_19, https://doi.org/10.1007/978-3-030-23204-7_19
25. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. ArXiv (2019)
26. Lui, M., Baldwin, T.: langid.py: An off-the-shelf language identification tool. In: Proceedings of the ACL 2012 System Demonstrations. pp. 25–30. Association for Computational Linguistics, Jeju Island, Korea (2012), https://aclanthology.org/P12-3005
27. Manakul, P., Liusie, A., Gales, M.J.F.: Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models (2023)
28. McCallum, A.K.: Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu (2002)
29. Napanoy, J., Gayagay, G., Tuazon, J.: Difficulties encountered by pre-service teachers: Basis of a pre-service training program. Universal Journal of Educational Research **9**, 342–349 (02 2021). https://doi.org/10.13189/ujer.2021.090210
30. OpenAI: Gpt-4 technical report (2023)
31. Plutchik, R.: A psychoevolutionary theory of emotions. Social Science Information **21**(4-5), 529–553 (1982). https://doi.org/10.1177/053901882021004003
32. Schmid, H., Laws, F.: Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging. In: Proceedings of the 22nd International Conference on Computational Linguistics - COLING '08. Association for Computational Linguistics, Morristown, NJ, USA (2008)
33. Shashkov, A., Gold, R., Hemberg, E., Kong, B., Bell, A., O'Reilly, U.M.: Analyzing student reflection sentiments and problem-solving procedures in moocs. In: Proceedings of the Eighth ACM Conference on Learning @ Scale. p. 247–250. L@S '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3430895.3460150
34. Sidarenka, U.: PotTS: The Potsdam Twitter sentiment corpus. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). pp. 1133–1141. European Language Resources Association (ELRA), Portorož, Slovenia (May 2016), https://aclanthology.org/L16-1181
35. Solopova, V., Popescu, O.I., Chikobava, M., Romeike, R., Landgraf, T., Benzmüller, C.: A German corpus of reflective sentences. In: Proceedings of the 18th International Conference on Natural Language Processing (ICON). pp. 593–600. NLP Association of India (NLPAI), National Institute of Technology Silchar, Silchar, India (Dec 2021), https://aclanthology.org/2021.icon-main.72
36. Ullmann, T.: Automated analysis of reflection in writing: Validating machine learning approaches. International Journal of Artificial Intelligence in Education **29** (02 2019). https://doi.org/10.1007/s40593-019-00174-2
37. Wojatzki, M., Ruppert, E., Holschneider, S., Zesch, T., Biemann, C.: GermEval 2017: Shared task on aspect-based sentiment in social media customer feedback. In: Proceedings of the GermEval 2017 – Shared Task on Aspect-based Sentiment in Social Media Customer Feedback. pp. 1–12. Berlin, Germany (2017)
38. Wulff, P., Buschhüter, D., Westphal, A., Nowak, A.I., Becker, L., Robalino, H., Stede, M., Borowski, A.: Computer-based classification of preservice physics teachers' written reflections. Journal of Science Education and Technology **30**, 1–15 (2020)

Appendixes

Table 1. Metrics mentioned in the paper.

Metric	Definition
F1-score	A harmonic mean of the precision and recall calculated per class. Can range from 0 to 1.
F1-score macro	The metric is computed independently for each class and then the average is taken.
F1-score micro	The metric aggregates the contributions of all classes to compute the average metric.
Cohen's kappa	The metric is used to measure inter-annotator reliability for categorical items. 0.41–0.60 is interpreted as moderate agreement, 0.61–0.80 as substantial, and 0.81–1.00 as perfect agreement.
QWK	Quadratic Weighted Kappa measures the agreement between two outcomes ranging from -1 (complete disagreement) to 1 (complete agreement).
Hamming score	The metric is often used for multi-label classification calculating the fraction of wrong labels to the total number of labels. The values higher than 0.9 are excellent scores, higher than 0.7 are good scores, and lower than 0.7 may be considered poor.

Table 2. Topics clusters from Bertopic.

Clustering 1	Clustering 2
Lectures and editing	Teamwork and Tasks
Classroom Management	Teacher, school, teaching
Pedagogy and Educational Diagnostics	Algorithms, Computer Science, Digital Technology
Reading and Literature	Self-promotion
Conflict Analysis	Music
Feedback	Math and numeracy
Your Subject Area	Science and Experiments
Diagnostics and diagnostic procedures	
Intervention measure	
Motivation	
Portfolio	
Lecture material and video	
Psychology	

Table 3. Emotion detection labels.

Emotions & Feelings
information
annoyance
appreciation
disapproval/critique
interest
anticipation
excitement
challenged
confidence
disappointment
insecurity
motivation
optimism
responsibility
satisfaction
surprise
uncertainty
wariness

Table 4. Defenitions of reflective labels.

Level	Defenition
Description	It is the lowest level, where the person only describes the circumstances and may include an evaluation of their own feelings.
Reflective description	Here one's own perspective analysis and superficial justifications are present.
Dialogical Reflection	It includes analysis of various perspectives as if in form of an internal dialogue with oneself.
Transformative Reflection	It should include the plan for the next steps or what one would do next time in such a situation.
Critical Reflection	The highest level of reflection encompasses a wider context (social, political, historical).