Pre-selecting Text Snippets to provide formative Feedback in Online Learning

Sylvio Rüdian
Humboldt-Universität zu Berlin
ruediasy@informatik.hu-berlin.de

Clara Schumacher
Humboldt-Universität zu Berlin
clara.schumacher@hu-berlin.de

Jakub Kuzilek
Humboldt-Universität zu Berlin
jakub.kuzilek@hu-berlin.de

Niels Pinkwart
German Research Center for Artificial Intelligence
niels.pinkwart@dfki.de

ABSTRACT

In this paper, a proof of concept is shown to generate formative textual feedback in an online course. The concept is designed to be suitable for teachers with low technical skill levels. As state-of-the-art technology still does not provide high-quality results, the teacher is always held in the loop as the domain expert who is supported by a tool, and not replaced. The paper presents results of our proposed approach for semi-automatic feedback generation using a real-world university seminar, where students create sample micro-learning units as online courses, for which they get feedback for. A supervised machine learning approach is trained based on learner submissions features, and the feedback, that was chosen by teachers in former submissions. The results are promising.

Keywords

Formative Feedback, Online Learning, Teacher Support, Prediction.

1. INTRODUCTION

Feedback is considered essential for supporting successful learning processes and outcomes [1]. Feedback can be defined as „information provided by an agent (e.g., teacher, peer, book, parent, self, experience) regarding aspects of one's performance or understanding" [1]. However, the timely provision of elaborated individual feedback is limited due to large student cohorts and limited resources in higher education. The lack of resources results in predominant use of summative assessments (and feedback) [2], which are often used at the end of a learning unit or course for grading and certification purposes if predefined objectives are met [3]. Due to heterogeneous students, the provision of individual support is even more relevant. Therefore, formative assessments aiming at providing students feedback on their performance or next learning steps is crucial. Instead of being distinct concepts, the functions of formative and summative assessments are on a continuum as such that the engagement with assessment tasks or the potential feedback can result in a change of learners' behavior [4]. In sum, elaborated feedback offering information on task-, process- and self-regulation level has been found to be most effective for learning success [5]. This includes an understanding of the learning goals that need to be achieved („Where the learner is going"), assessing the evidence of learning („Where the learner is right now"), and the provision of feedback on how to achieve the designated learning goals [6]. However, even with the increased use of digital learning environments and methods such as learning analytics the provision of informative feedback at scale is challenging [7] and time-consuming. Due to the need of extra resources, formative feedback is often not provided at all, or solely on the correctness, in the form of sample solutions or short paragraphs. The paper aims to support teachers in the process of giving textual feedback.

S. Rüdian, C. Schumacher, J. Kužílek, and N. Pinkwart. Pre-selecting text snippets to provide formative feedback in online learning. In M. Feng, T. Käser, and P. Talukdar, editors, *Proceedings of the 16th International Conference on Educational Data Mining*, pages 430–433, Bengaluru, India, July 2023. International Educational Data Mining Society.

https://doi.org/10.5281/zenodo.8115748

2. RELATED WORK

Automated feedback can be characterized based on several properties [8]: a) the adaptiveness of the feedback; b) its timing; c) learners' control over the feedback: and d) the purpose of the feedback. Automated feedback can for example be not adaptive at all, dependent on students' solution to a task or also on their characteristics and learning behavior. Timing of the automated feedback can be immediate after the action, upon request or at the end of a task. The feedback provision might further be controlled by the learner for example with regards to the amount and frequency of feedback, the timing, its appearance. The need for control has also been brought up by studies investigating students' preferences of automated interventions (e.g. [9]). The purpose of the feedback refers to simple corrective feedback, suggestion of future actions, additional information, or motivational feedback [8]. Despite the examination of computer-generated feedback for decades; still the creation of highly informative feedback is very complex, where machines can be supportive, but do not replace teachers [10]. As texts created by learners are manifold and diverse it is hard to evaluate them automatically [11]. Due to the low quality of computer-generated feedback, its use can lead to high frustration [12]. For example, available state-of-the-art automatic writing evaluation tools, such as proofreading tools to detect mistakes in submissions of language learners, do not meet teachers' expectations [13]. Hence, the teacher is vital for the provision of feedback. Thus, instead of providing computer-generated feedback to learners directly, a teacher-in-the-loop approach is of high importance. Therefore, the process to create feedback must be intuitive without the need for complex adjustments.

In the domain of education, decisions coming from computer-generated feedback tools must be explainable. This is a key component of the trusted learning analytics approach (TLA) [14]. One possible solution is the tool OnTask [15], which can principally be used to generate texts based on pre-defined text snippets and rules that use trace data. Based on such rules, decisions can be justified and explained. If for example, the learner submits a text and the tool recognizes that the learner skipped watching a related learning video, which is implemented as a rule, then feedback is given using the snippet with the advice to have a deeper look at the learning

material. However, it is essential to educate teachers so that they get an understanding of the versatility of such software. Teachers must have scenarios in mind, which must be implemented in rules. From the practical perspective, this is a pitfall as teachers want to focus on their domain to create learning material and not on scenarios that possibly can exist [12]. Hence, feedback is mainly limited to tasks, where feedback can be predefined. For multiple-choice questions, feedback can be given if the correct choices are selected, but also for incorrect selections, respectively. Considering textual submissions, the state-of-the-art Moodle, and H5P versions allow searching for specific keywords. If they are missing in the text, feedback can be provided. Nevertheless, such feedback assumes that the learner uses concrete vocabulary (or synonyms, that are predefined by the teacher). If they use other words or descriptions, they still get the same feedback as others, which can lead to frustration. If learner texts are aimed to be evaluated on an individual level automatically, the topic of automatic essay scoring (AES) emerges. There, texts are scored, intending to compare learners' results. Most AES systems have in common, that they need to be trained with a large sample size with annotated texts and they extract a huge number of linguistic features [16]. Exemplarily, the AES „IntelliMetric" [17] extracts over 300 features, ranging from conceptual, and structural features to rhetorical attributes [18]. First, the approach examines cohesiveness and consistency. Then, the scope of the content is analyzed, followed by an evaluation of text structure, and transitional fluency. Then, sentence structure is investigated, using sentence complexity with readability metrics, and syntactic variety. Finally, mechanics and conventions are analyzed, to test whether the text is in line with standard American English (spelling, grammar, etc.) [16]. However, most tools are not open-source and rely on financial benefits. Thus, their application is limited to institutes, which have the budget to spend.

3. FRAMEWORK

Following Deeva et al. [8], automated feedback can be expert-driven as in the rule-based systems (e.g., OnTask), or data-driven considering student data using algorithmic approaches or a combination of both. In the proposed framework, the importance of the teacher in the loop is emphasized. The idea of having the teacher-in-the-loop is extended by Rüdian et al. [19] to connect learner submissions with feedback by exploring derived NLP features and its relation to feedback given in concrete contexts. To the best of our knowledge, this concept has not been applied to a real-world online course setting. Thus, we focus on the research question of whether there is a set of NLP features (extracted from learner submissions), that are predictive to auto-select ratings, which were previously selected by teachers.

The approach proposes a teacher-in-the-loop approach that is based on pre-defined text snippets to provide feedback on task-level. Such snippets can be extracted from already given feedback texts or best practices in the literature. Text snippets must meet the condition to be related to a scale (e. g., Likert scale, binary (yes/no) scale). In the training process, teachers create feedback by selecting pre-defined text snippets. The idea of using such snippets is not new, but a helpful step for teachers to reduce the required time to create feedback [19]. Then, snippets are stored including the rating on the scale, e. g. whether a learner correctly applied a concept, or not. NLP features are extracted from user artifacts (e. g. textual submissions). Features can be based on sentiment analysis, word-sense disambiguation, argument mining, or others [18]. Such features are then used to train a supervised machine learning approach, aiming to predict ratings on evaluation criteria. Explainable methods such as the Naïve Bayes classifier [20] are favored to follow the TLA approach. For all labels that can be predicted with acceptable accuracy, a model is stored. Then, for new learner artifacts of the same task, ratings can be predicted. The teacher gets those predictions so that related text snippets are automatically pre-selected when the teacher aims to create feedback. Based on those selections, a final feedback text is generated. Besides, a reinforcement learning approach is used. The teacher can change pre-selections. Thus, new training data are continuously created to train the model with more data to become more generalizable. Also, the student can evaluate feedback to obtain a critical view of its applicability. The main idea is to separate teachers from the machine learning approach, that runs in the background.

4. STUDY DESIGN

In a university seminar, students have the task to design a micro-learning online course (~15-25 min) covering a topic of their choice. Students create courses in a Moodle instance. 33 courses are created. They receive feedback from a tutor who uses a form of 28 evaluation criteria and selects whether the criteria are fulfilled. Selections must be rated on a Likert (5=totally agree to 1=totally disagree) or binary scale (the latter is used for the case, where only two options exist). For binary options, also 5 (agree), and 1 (do not agree) are used. Feedback criteria are based on literature research to rate the quality of an online course. In detail, clearness, instructions, and learning materials are rated, whether appropriate feedback is given [23], learning goals and expectations are included [24], a target group is defined, and whether the course content is appropriate for those learners [25]. Further, it is rated whether designed tasks have an appropriate difficulty level and whether final tests are suitable to evaluate knowledge gain [26], and, of course, the correctness of the created learning material.

The tutor uses the system to generate a feedback text, based on his/her selections, which is the standard process in this setting to provide feedback. The automatically generated text can be changed or enhanced by the tutor. However, as to date, further text adjustments are only used to a negligible amount by the tutors; this will be investigated at a later stage in more detail and is not covered in this paper. Selected feedback options are stored for each course that is submitted by students. Those courses are the artifacts and build the base for the data set. Thus, the courses are used as the input variables and the aim is to pre-select the rating on the evaluation criteria, that are used to generate the textual feedback.

Then, an experimental analysis is done to examine the predictability of the items. Textual features must first be extracted from all courses. To do that, courses are transferred to a CSV file using Moodle backups of the courses, and from that, the main information is extracted. Each line is related to an item of the course progression. The CSV file contains the item type (more detailed, whether H5P is used, a content page is created, or the Moodle quiz tool is used). It contains the header of the item, the content, and in case of interactive items (H5P/quiz), also questions including responses, correctness, and feedback. Based on that information, the course can principally be reconstructed. As a CSV file is created for each course, a transfer to a feature vector is required, containing the same number of features for each course, aiming to train a predictive model.

The following features are extracted and stored in a new CSV file:

(1) Number of items, including types (H5P, pages, Moodle quiz, videos),
(2) Text complexity metrics of the content, and questions (Flesh Reading Ease [25], or Gunning Fog Index [26]),

(3) Use of keywords in texts („target group", „references/literature"),
(4) Number of items, where feedback is given, namely feedback given on wrong, or correct responses, and overall feedback,
(5) Polarity and subjectivity of contents.

Before training an approach to make predictions, the distribution of selected options is analyzed to detect highly imbalanced options. Due to the limited number of courses (33), ratings on a Likert scale are not well balanced. To be still able to predict those ratings, ratings are transformed to a binary scale on indicating that the criterion is met (5, criterion passed) and one representing all other values (1-4; at-risk, criterion failed). Criteria, that are still imbalanced (like all students fulfilled them), are filtered out as it is not worth examining predictability due to the limited dataset. To give an example: All students described the learning goal in their course. Thus, there is no low rating for the criterion, so it can be ignored. For the remaining ones, distributions are explored to see whether there is a remarkable difference, considering all features separately. Those, where a difference can be seen, are selected for the proof-of-concept. Then, a Naïve Bayes model is trained, as it is easy to interpret probabilities, which fulfills the TLA condition. Besides, it can easily be extended to a multidimensional problem and the resulting trained model can easily be implemented by using web technologies without the necessity of deploying complex computational power. The resulting predictions are evaluated using 5-fold cross-validation.

5. RESULTS

Based on the initial analysis of the distributions of binary ratings for all criteria and all features, the target variable is chosen on whether *the learning goal is covered by a final test*. There have been 11 failed and 22 passed cases. For the selected target variable, the corresponding extracted features distributions with a focus on the target class has been visually analyzed and the most promising 4 features have been selected for the initial proof-of-concept. Selected features are: the number of Moodle quiz items; the number of feedbacks given for correct responses; question readability grades ARI (Automated Readability Index); and question readability grades for Flesch Reading Ease (FRE). Remaining features are excluded. The corresponding distributions are depicted in Figure 1.

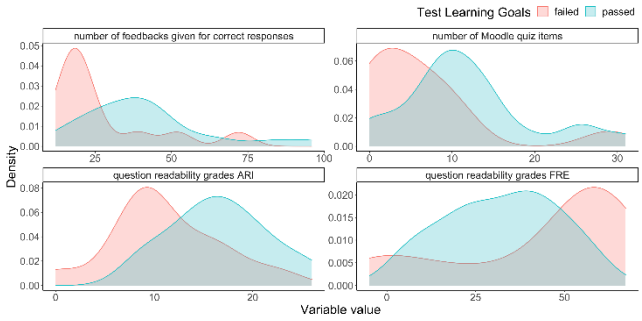

Figure 1: Distributions of binary classes for four features.

Both cases (passed vs. at-risk/failure) are plotted with red, and blue colors. For selected features, differences in distributions can be seen. This is a good sign, as those features split the dataset by the binary ratings in general. Following the selected features, the Naïve Bayes model with Gaussian kernel is trained using 10-fold cross-validation. The error is estimated using 5-fold cross-validation covering the complete data set. Thus, the process of error estimation and model training is as following: data is divided into 5 folds using stratified sampling without replacement and in 5 steps, the model is trained using 4 folds of input data via 10-fold cross-validation and the error is estimated with the remaining 1 data fold.

To simplify the model for the deployment, we explored 5 different scenarios: using each feature separately (4 scenarios) and using selected features together to train the model. Table 1 reports the results using mean values accuracy (Acc), precision (P), and recall (R) in 5 rounds of cross-validation. P and R are computed for both classes (passed and at-risk/failed) to understand their predictive power (guessing would be .5 for the binary option). As visible, the feature of the „number of given feedback on correct responses" outperforms scenario 5, where all features are used together.

Table 1. Results for four features.

Feature	Acc	P	R	P	R
		passed		failed	
Number items quiz	.68	.74	.78	.63	.46
Number of given feedback on correct responses	.75	.87	.77	.63	.73
Question readability ARI	.76	.80	.86	.81	.56
Question readability FRE	.73	.75	.90	.60	.43
All features	.71	.77	.77	.63	.60

6. DISCUSSION

Compared to a pre-defined rule-based approach the proposed approach allows to provide more fine-grained feedback and dynamic support. Furthermore, it aims at enhancing teachers' practices and reducing their workload for providing highly informative feedback on text artifacts. Thus, the approach considers the limited resources in higher education for providing formative feedback but still enables learners to derive appropriate future learning activities. Due to complexity of algorithms and their limitedness of providing actionable outcomes a major concern in educational settings is the limited acceptance of the stakeholders. This might be avoided by the simplicity of the proposed approach that enables teachers to create feedback without the need for abstract technical skills plus by being grounded in the idea of TLA of having the human in the loop of an explainable approach.

This proof-of-concept is limited as only data of students that agreed to share their data for this research were analyzed resulting in 33 submissions which might have led to biases. This calls for future research with larger data sets.

From a statistical perspective, computational complexity involves the estimation of the Gaussian distribution during the model training and then, it compares two posterior probabilities. In the final model, only two equations (for estimation of the probabilities) and their comparisons are computed. We also limited ourselves to the most promising features and selected one criterion for which the concept is working well. The training step is required for each criterion, requiring to create 28 separate models. Thus, in future a more refined approach which can do the estimation at once (for example by mapping the separate criteria to another dimension, where one number reflects unique criteria combination) will be explored. The restriction to binary cases is necessary to simplify the small dataset, in future work either the one-vs.-rest/one-vs.-one approach or a regression model for proper estimation of the scale values will be investigated. However, if this is aimed to be examined, the dataset must be extended with further samples.

The accuracy of using all features suggests, that the model tends to overfit with higher dimensions. Using one feature leads to better

results, less computational complexity, and ease of use. Values of precision and recall are better in general for the case of passed class. This is probably because even with the selected binary target value, classes are imbalanced. Thus, the trained model prefers positive cases due to a better fit of the distribution. Further, we used a handful of NLP features, extracted from learner submissions. Exploring more linguistic features is of high interest to explore its predictive power. However, the approach needs to be enhanced further to support the teacher more as still the human needs to validate the feedback which might be also time-consuming. Furthermore, also students' behavioral data should be considered, for example, to determine the timing of the feedback as well as its properties (e.g., provided by a system or an e-mail of the tutor (see further [27]). As the uptake and actual use of feedback by the students is key for its effectiveness [28], their perceptions of the feedback [7] as well as their actions taken need to be investigated in more detail. Using experimental study designs the impact of the feedback on students' learning processes and outcomes will be investigated in detail. In sum, the proof-of-concept is promising, and predictions have been working in the concrete setting.

7. ACKNOWLEDGEMENTS

This work was supported by the German Federal Ministry of Education and Research (BMBF), grant number 16DHBKI045.

8. REFERENCES

[1] J. Hattie and H. Timperley, "The power of feedback" in *Review of educational research 77(1)*, 2007, pp. 81-112.

[2] J. Broadbent, E. Panadero and D. Boud, "Implementing summative assessment with a formative flavour: a case study in a large class" in *Assessment and Evaluation in Higher Education 43(2)*, 2017, pp. 307-322.

[3] V. J. Shute and B. J. Becker, "Prelude: Assessment for the 21st century" in *V. J. Shute & B. J. Becker (Eds.), Innovative Assessment for the 21st Century. Supporting Educational Needs* , Springer, 2010, pp. 1-11.

[4] P. Black and D. Wiliam, "Classroom assessment and pedagogy" in *Assessment in Education: Principles, Policy & Practice, 25(6)*, 2018, pp. 551-575.

[5] B. Wisniewski, K. Zierer and J. Hattie, "The power of feedback revisited: A meta-analysis of educational feedback research" in *Frontiers in Psychology 10*, 2020, p. Article: 3087.

[6] D. Wiliam and M. Thompson, "Integrating assessment with learning: What will it take to make it work" in *C. A. Dwyer (Ed.), The Future of Assessment. Shaping Teaching and Learning*, Lawrence Erlbaum Associates, 2008.

[7] L.-A. Lim, S. Dawson, D. Gašević, S. Joksimović, A. Fudge, A. Pardo and S. Gentili, "Students' sense-making of personalized feedback based on learning analytics" in *Australasian Journal of Educational Technology 36(6)*, 2020, pp. 15-33.

[8] G. Deeva, D. Bogdanova, E. Serral, M. Snoeck and J. De Weerdt, "A review of automated feedback systems for learners: Classification framework, challenges and opportunities" in *Computers & Education (162), 104094.*, 2021.

[9] C. Schumacher and D. Ifenthaler, "Features students really expect from learning analytics" in *Computers in Human Behavior (78)*, 2018, pp. 397-407.

[10] E. C.-F. Chen and W.-Y. E. C. Cheng, "Beyond the design of automated writing evaluation: Pedagogical practices and perceived learning effectiveness in EFL writing classes" in *Language Learning & Technology 12.2*, 2008, pp. 94-112.

[11] T. K. Landauer, "Automated scoring and annotation of essays with the Intelligent Essay Assessor" in *AES: A cross-disciplinary perspective*, 2003, p. 87–113.

[12] P. Ware, "Computer-generated feedback on student writing" in *Tesol Quarterly 45.4*, 2011, pp. 769-774.

[13] S. Rüdian, M. Dittmeyer and N. Pinkwart, "Challenges of using auto-correction tools for language learning" in *LAK22*, 2022, pp. 426-431.

[14] J. Hansen, C. Rensing, O. Herrmann and H. Drachsler, "Verhaltenskodex für Trusted Learning Analytics. Version 1.0. Entwurf für die hessischen Hochschulen" in *Innovationsforum Trusted Learning Analytics*, 2020.

[15] A. Pardo, K. Bartimote, S. B. Shum, S. Dawson, J. Gao, D. Gašević and L. Vigentini, "OnTask: Delivering data-informed, personalized learning support actions" 2018, pp. 235-249.

[16] S. Dikli, "Automated Essay Scoring" in *Turkish Online Journal of Distance Education-TOJDE (7)*, 2006.

[17] L. M. Rudner, V. Garcia and C. Welch, "An evaluation of IntelliMetric™ essay scoring system" in *The Journal of Technology, Learning and Assessment 4.4*, 2006.

[18] intellimetric, "How it Works" Vantage Labs, 17 06 2017. [Online]. Available: https://www.intellimetric.com/how-it-works. [Accessed 05 01 2023].

[19] S. Rüdian. C. Schumacher, N. Pinkwart: "Computer-Generated formative Feedback using pre-selected Text Snippets" in *LAK*, 2023, pp. 129-131.

[20] K. P. Murphy, "Machine learning: a probabilistic perspective", MIT press, 2012.

[21] Q. Matters, "K-12 Rubric Workbook Standards for Course Design (Fifth Edition)", Annapolis, MD: Maryland Online, 2019.

[22] D. Xu, Q. Li and X. Zhou, "Online course quality rubric: a tool box" in *Online Learning Research Center*, 2020.

[23] S. Baldwin, Y. H. Ching and Y. C. Hsu, "Online course design in higher education: A review of national and statewide evaluation instruments" in *TechTrends 62(1)*, 2018, pp. 46-57.

[24] S. D. Achtemeier, L. V. Morris and C. L. Finnegan, "Considerations for developing evaluations of online courses" in *Journal of Asynchronous Learning Networks 7.1*, University of Georgia, 2003, pp. 1-13.

[25] R. Flesch, "A new readability yardstick" in *Journal of applied psychology*, *32*(3), 221, 1948.

[26] R. Gunning, "The Technique of Clear Writing", McGraw-Hill, 1952.

[27] C. Schumacher, D. Ifenthaler, "Investigating students' perceptions of system- vs. teacher-based learning analytics feedback", 2023.

[28] N. E. Winstone, R. A. Nash, M. Parker and J. Rowntree, "Supporting learners' agentic engagement with feed-back: A systematic review and a taxonomy of reciepience processes" in *Educational Psychologist 52(1)*, 2017, pp. 17-37.