Lecture Notes on Data Engineering and Communications Technologies 190

Tim Schlippe
Eric C. K. Cheng
Tianchong Wang *Editors*

Artificial Intelligence in Education Technologies: New Development and Innovative Practices

Proceedings of 2023 4th International Conference on Artificial Intelligence in Education Technology

Springer

Lecture Notes on Data Engineering and Communications Technologies 190

Series Editor

Fatos Xhafa, *Technical University of Catalonia, Barcelona, Spain*

The aim of the book series is to present cutting edge engineering approaches to data technologies and communications. It will publish latest advances on the engineering task of building and deploying distributed, scalable and reliable data infrastructures and communication systems.

The series will have a prominent applied focus on data technologies and communications with aim to promote the bridging from fundamental research on data science and networking to data engineering and communications that lead to industry products, business knowledge and standardisation.

Indexed by SCOPUS, INSPEC, EI Compendex.

All books published in the series are submitted for consideration in Web of Science.

Tim Schlippe · Eric C. K. Cheng ·
Tianchong Wang
Editors

Artificial Intelligence in Education Technologies: New Development and Innovative Practices

Proceedings of 2023 4th International Conference on Artificial Intelligence in Education Technology

Springer

Editors
Tim Schlippe
IU International University of Applied Sciences
Erfurt, Germany

Tianchong Wang
Swinburne University of Technology
Melbourne, VIC, Australia

Eric C. K. Cheng
The Education University of Hong Kong
Hong Kong S.A.R., China

ISSN 2367-4512 ISSN 2367-4520 (electronic)
Lecture Notes on Data Engineering and Communications Technologies
ISBN 978-981-99-7946-2 ISBN 978-981-99-7947-9 (eBook)
https://doi.org/10.1007/978-981-99-7947-9

© The Editor(s) (if applicable) and The Author(s), under exclusive license
to Springer Nature Singapore Pte Ltd. 2023, corrected publication 2023, 2024

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Paper in this product is recyclable.

Preface

The 2023 4th International Conference on Artificial Intelligence in Education Technology (AIET 2023) was held in Berlin, Germany, during June 30 – July 2, 2023. The conference provides a unique platform for international and domestic researchers in the field of artificial intelligence in education technology to share their latest research results and findings.

Over the years, the conference has continued to grow and develop with the help and support of many scholars, and experience has been accumulated year by year. In 2023, 56 papers were submitted, and we were pleased to have accepted 24 papers from the total submissions in this proceeding. All of the papers were subjected to double-blind peer review by the program technical committee members and international reviewers.

Researchers, engineers as well as industrial professionals from all over the world presented their research results at the conference. The hybrid conference lasted three days, including online test and onsite check-in/materials collection on the first day. The following two days included the keynote speeches, invited speeches, online/onsite oral presentation sessions, best presentations awards and best student paper award. The proceedings were divided into four chapters depending on the specific topics: educational data mining and learning analysis, text analysis and natural language processing for topics related to education, design and evaluation of educational information systems, instruction of artificial intelligence and artificial intelligence integrated education. It was a golden opportunity for the students, researchers and engineers to interact with the experts and specialists to get their advice or consultation on artificial intelligence in education technology.

I would like to thank all the authors who have contributed to this volume and also to the organizing committees, reviewers, speakers, sponsors and all the conference participants for their support to AIET 2023!

With my warmest regards,

Tim Schlippe
Conference General Chair

Conference Committee

General Chairs

Sven Schütt	IU International University of Applied Sciences, Germany
Tim Schlippe	IU International University of Applied Sciences, Germany

Conference Co-chair

Mario Barajas Frutos	University of Barcelona, Spain

Program Chairs

Eric C. K. Cheng	The Education University of Hong Kong, China
Yanjiao Chen	Zhejiang University, China

Program Co-chairs

Grigorios N. Beligiannis	University of Patras, Greece
Yongbin Zhang	Beijing Institute of Graphic Communication, China

Publication Chair

Dariusz Jacek Jakóbczak	Technical University of Koszalin, Poland

Local Chair

Matthias Wölfel	Karlsruhe University of Applied Sciences, Germany

Technical Program Committees

Yu-Mei Wang	University of Alabama at Birmingham, USA
Norma Binti Alias	Technology University of Malaysia, Malaysia
Loc Nguyen	International Engineering and Technology Institute (IETI), China
Francesco Flammini	Mälardalen University, Sweden
Mirela Müller	University of Split, Croatia, Croatia
Kuldeep Rawat	Elizabeth City State University, USA
Ke Deng	RMIT University, Australia
Min Lun (Alan) Wu	Ohio University, USA
Lei Niu	Central China Normal University, China
Hector Rafael Morano Okuno	Tecnologico de Monterrey, School of Engineering and Science, Mexico
Michael Losavio	University of Louisville, Louisville, Kentucky, USA
Simon Baradziej	UiT The Arctic University of Norway, Norway
Adrian Groza	Technical University of Cluj-Napoca, Romania
Cynthia Breazeal	MIT Media Lab, USA
Dieter Uckelmann	University of Applied Sciences, Germany
Hyun Sook Yi	Konkuk University, South Korea
Sule Tasli Pektas	OSTIM Technical University, Turkey
Kristina Schaaff	IU International University of Applied Sciences, Germany
Eryka Wilson	ReadyAI, Keiser University, USA
Deepti Mishra	Norwegian Institute of Science and Technology (NTNU), Norway
Jia Chen	Central China Normal University, China
Lu Han	Central University of Finance and Economics, China
Huichuan Dai	Guangdong University of Science and Technology, China

Contents

Text Analysis and Natural Language Processing for Topics Related to Education

Design and Evaluation of Educational Information Systems

Educational Data Mining and Learning Analysis

A Qualitative Evaluation of an AI-Based Study Progress Forecast

Sudarshan Kamath Barkur(✉), Mascha-Lea Fersch(✉), Sophie Henne(✉), Sigurd Schacht(✉), and Betiel Woldai(✉)

University of Applied Science Ansbach, Ansbach, Germany
{s.kamath-barkur,mascha-lea.fersch,sigurd.schacht, b.woldai}@hs-ansbach.de, sophie.henne@web.de

Abstract. This paper presents the development and evaluation of a first prototype of an ai-based study progress forecast. This service is integrated within a conversational agent and can be used by students to show them their current study progress. First, implications for the set-up of a forecast application from the literature are described. Based on the requirements identified in the literature and from the project itself, a lightweight formula was created that enables calculating the remaining study time. In order to assess preliminary feasibility and perception of the model prototype, a qualitative focus group discussion was conducted with five participants. Overall, the study progress forecast was well received by the participants, especially the offer itself as well as the promptness of the service were highlighted.

Keywords: higher education · study progress forecast · digital study assistant

1 Introduction

According to a study by the DZHW, around a quarter of all Bachelor's students drop out of their studies at a university [1]. Reasons for this development are usually complex, but individual performance problems, financial problems during studies, and lack of motivation can be cited as frequently occurring factors [2]. A high dropout rate imposes various challenges to universities in the form of, among other things, mis-investment or even a reduction in the allocation of financial resources by the state [3].

It is therefore in the interest of universities to keep the dropout rate as low as possible. Various instruments and preventive measures exist for this purpose, such as mandatory motivational letters, offering introductory courses or a mentoring program to promote social integration and identification with the university [3]. Another measure is introducing an early warning system to detect students at risk of dropping out. Based on relevant data, these systems provide an automated feedback in scenarios where intervention appears necessary to avoid the deregistration of a student [4].

In a research project at the Ansbach University of Applied Science, the development of an early warning system based on artificial intelligence is currently being planned. The service enables a voluntary determination of the study's progress. The forecast should

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023
T. Schlippe et al. (Eds.): AIET 2023, LNDECT 190, pp. 3–13, 2023.
https://doi.org/10.1007/978-981-99-7947-9_1

realistically show students' potential risks in their study progress. Such developmental analysis could help reduce failure rates even in the short term. As a short- to medium-term effect, the planned innovation should increase students' information level, satisfaction, and motivation [4]. In the long term, it should contribute to appropriate study times and the reduction of dropout rates.

In the following paper, the design and development of the service will be presented. In order to lay the foundation for the application design, related work identified from a literature research will be outlined in Sect. 2. Based on these findings the application design will be described in Sect. 3. To ensure functionality and quality of the service, Sect. 4 presents the methodology and results of a qualitative focus group. Finally, insights from the development process and the focus group are discussed in Sect. 5, followed by a conclusion in Sect. 6.

2 Theoretical Background

[4] define an early warning system as a model that uses threshold values based on suitable key figures, which trigger an "alarm" if they are exceeded or not reached. Further, this signal is linked to information, advice and support services for students.

Der Einsatz von Early Warning Systemen an Hochschulen wird insbesondere deshalb untersucht, da Studienabbrüche weitreichende ökonomische Auswirkungen sowohl auf Hochschulen, auf Studierende als auch auf die Gesellschaft verursachen.

Depending on the university, the selection of the data basis to be used such as administrative student or examination data as well as suitable indicators and associated threshold varies [5]. Indicators mentioned in the literature include credit points achieved in relation to planned credit points, number of examination attempts, number of withdrawals from examination attempts and grade point average [4]. The type of feedback to students also varies from pure information about a possible risk in the study progress to concrete hints about support offers or invitations to counseling sessions. A technical infrastructure, e.g., software for data collection and evaluation, is necessary to implement such a system.

Already existing implementations can be used to further describe the concept of early warning systems at universities. Therefore a research was conducted to identify examples within the German higher education environment. Research projects from the Federal Ministry of Education and Research funding line with the topic "Studienerfolg und Studienabbruch I & II" were considered as a starting point [6]. After reviewing the funded projects, it can be stated that the results are predominantly oriented as services for colleges and universities, i.e., prediction models were developed for identifying students at risk of dropping out, which serve them as clues for possible interventions as well as support services. The DMPS research project at the University of Duisburg-Essen, for example, focused on the investigating and developing machine learning models that universities can use to identify students at risk of dropping out [7]. A random forest approach was proposed modeling the dropout decision as binary classification problem using data that considers aspects of the entire study program such as grade of university entrance qualification, satisfaction, and examination success [5]. The model's ability to make accurate predictions could be improved by including information from the early study phase along with pre-study characteristics [5]. The early warning system

"FragSte" developed at the University of Wuppertal uses administrative student data such as demographic variables, school education and academic performance to derive clues for possible interventions as well as support services based on the results [8]. However, the study highlights the trade-off between prediction accuracy and data requirements, which states that the accuracy of the prediction increases with the amount of data that needs to be considered. Specifically, in the first two semesters, a significant amount of data is needed to make accurate predictions, while in later semesters, the predictions become more precise due to the availability of more information about academic performance [8]. At the University of Freiburg, a study progress analysis system is used as a quality management tool to improve the quality of teaching [9]. The method used in the study involves combining data from the student administration with feedback from students on teaching events. As a result of the experiment, a significant correlation was discovered between the rate of students dropping out in their first year and the level of student satisfaction [9].

Since it is planned to establish a self-service for students to enable a study progress forecast, more research was conducted to identify similar approaches. An early warning system designed for students was developed at Mittweida University of Applied Sciences, which provides an overview of the performance status and progress of one's own studies based on program data, subject and university semesters, ECTS credit status and examination results etc. via an app [10]. The goal is to link student performance indicators with individually tailored advisory and support services. With the tool "PASST?!", the TU Dresden also offers a system that, with the student's consent, analyzes potentially problematic developments such as the acquisition of few ECTS points over a more extended period or the repeated failure to pass an examination and informs the student via e-mail about support offers [11].

3 Application Design and Architecture

Supported by the results identified from the research, requirements for the study progress forecast could be collected. As mentioned in Sect. 1, this idea is part of a research project at Ansbach university. The aim of the project is to develop a digital intelligent assistant for studying and teaching. Currently, an AI-based chatbot already exists, which is integrated into the website of the university and provides students with answers about their everyday study life or learning tips (see Fig. 1).

The bot's infrastructure is based on the RASA framework. The Rasa framework is an open-source machine learning tool that enables the development of chatbots and virtual assistants that can comprehend natural language, keep track of context, and handle complex conversations [12]. This framework was chosen being highly customizable and can be integrated with a variety of messaging platforms or other channels [12]. Thus, an environment that enables the integration of the feedback service already exists. However, this environment also frames the possibilities of the concept.

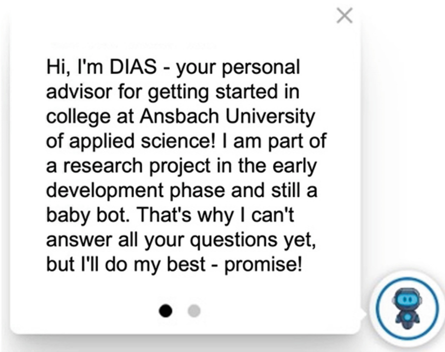

Fig. 1. The DIAS-Chatbot

The first decision was to specify how study progress should be defined. Currently, there is no integration of data bases containing any kind of student data planned for the assistant. That is why the study progress prediction must be based on different data. Moreover, requirements to protect sensible personal student data have to be followed. Since the existing assistant is a dialog-based system that can retrieve data, the chatbot intends to acquire relevant data in a conversation. To determine study progress, a lightweight formula was developed based on the examples for key figures from the literature. The remaining study time of a student should be calculated using the values achieved credit points so far and the current semester.

$$remainingStudyTime = \frac{necessaryCredits - currentCredits}{\frac{currentCredits}{currentSemester}}$$

Based on the entries, a calculation of whether enough credit points have been completed to finish the studies in the standard period of study is made. Thus, it can be illustrated whether the study progress is on schedule or whether negative deviations exist. In addition, the formula can be used to output the remaining credit points that are needed to catch up.

From a technical perspective, the module responsible for the calculations is connected to the RASA framework as an API. The function is started as soon as a user types in the keyword “study progress indicator” or selects it from the given options. Via RASA forms the relevant values for the formula “study type”, “actual Semester” and “actual credit points per semester” are collected from the user in a predefined format during a conversation (see Fig. 2).

The calculations are performed in the module and returned to RASA to be presented to the user as a response. The result is a text containing a recommendation depending on the case. Now, the entered data and calculations are not stored. The module’s functionality was tested manually with a human in the loop. The same applies to the connectivity between the RASA framework and the API.

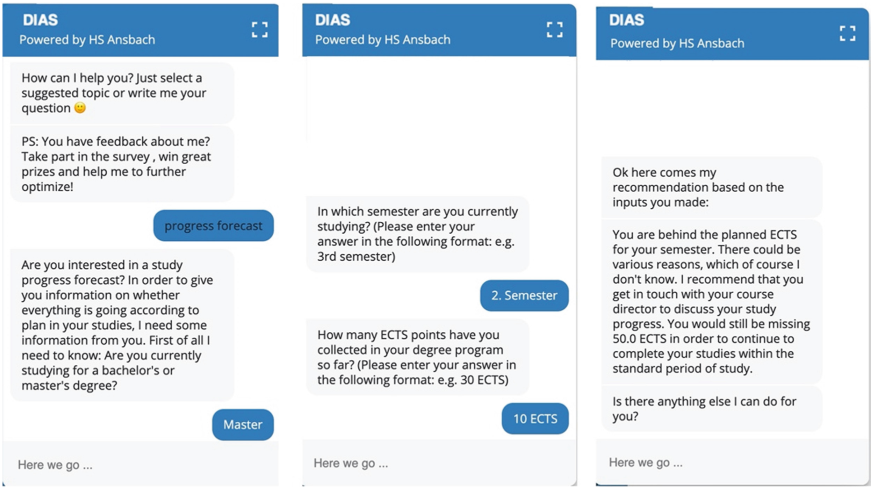

Fig. 2. Conversation Flow Study Progress Forecast

4 Evaluation

To evaluate the study forecast application from the perspective of future users, it was decided to have an initial prototype tested by students during the development process. The objective of this evaluation was to receive feedback on current features of the study forecast, but also ideas on how the function can be better adapted to the student's needs. For this purpose, a qualitative focus group was conducted with five participants. Moderated discussions are particularly suitable for requirement analysis, as they allow existing ideas and planned actions to be examined and discussed from different viewpoints, which can help identify potential problems and assess the feasibility of the proposed solutions [13]. This method has already been used to test further components of the digital assistant, such as the chat function [14].

4.1 Participants

Participants were recruited through a master's program at the Ansbach university of applied sciences. The focus group was finally conducted with one professor, three students and one scientific associate, all involved in the same study program. Among the participants were four females and one male. Due to the limited research funds, there was no reimbursement for the focus group participation.

4.2 Data Collection

The research team prepared a semi-structured interview manual to guide the focus group. This format was chosen to promote open discussions. The questions were partly created based on prior literature research of similar studies in the educational field [15]. An online conference via Zoom was chosen as the venue for the focus group. The session was moderated by two of the authors of this paper. The presentation and discussion were divided into two parts as two specifications of the analyzer were evaluated. In each part, the respective component was first presented and explained. Afterward participants had the opportunity to test the function themselves followed by a discussion. The discussions were recorded so that a transcript of the focus group could be created later (Table 1).

Table 1. Semi-structured Interview Guideline

Question Category	Explanation	Example Question
Scope & Content	Questions concerning the scope and content quality of the answers received in the forecasting feature	"Were the questions and answers always understandable for you?" Were you satisfied with the scope of the answers you received about the feature? Would you also be willing to give more information if anonymized?
Logic	Questions concerning the fundamental logic behind the model and the line of questioning	What other possibilities do you know/use to check your study progress? What other questions should the chatbot have asked you to be able to draw conclusions about your study progress? Do you think the prognosis, or the underlying logic adequately reflects your study progress?
User Experience	Questions dealing with usability, navigation, and the output format	How did you feel about navigating through the conversation? How did you feel about the chatbot as a channel to perform the study progress forecast?

4.3 Data Analysis

The data analysis of a focus group is usually based on a transcript [13]. For the interpretation, we used the qualitative content analysis method in two stages according to [16], since this procedure fits best to our research design. As a first step, the research team transcribed and reviewed the video recording several times. In a second step, the data were coded, categorized and after a final review adapted. The coding procedure was done using both an inductive and a deductive approach. As described previously the semi-structured interview guideline was based on three pre-defined categories. These categories were then adapted during the analysis, further classes were added and others were renamed. The coding process was reviewed by other team members to maximize objectivity.

4.4 Results

Overall, the study forecast feature was positively received by the participants. All students were in favor of using the application, especially with regard to the anonymity of the procedure.

In terms of functionality, the fast data output was highlighted. Only little information is required to calculate the remaining study time, moreover, the result is received within

Table 2. Results from the second part of the focus group, reflecting the testing of the study progress forecast. Texts have been translated from German.

Category and sub-category	Definition	Quote (example)
1. Future Use	All text passages that demonstrate the future interest in a continued use	"the function itself is totally awesome and I would absolutely use it." (Focusgroup2.2, paragraph 94)
2. Positive Aspects	All text passages that highlight the positive characteristics of the study progress forecast	"I personally thought it was great and I generally find it awesome that it can be done and that it works super fast." (Focusgroup2.2, paragraph 100)
3. Improvement Suggestions	All text passages that demonstrate improvement tips through the use of the study progress forecast	
3.1 User experience	All text passages that highlight improvement tips related to the user experience based on the input format and access through a chatbot	"What I did once was that I only entered the number of credit points during the input. (…) Maybe there could be a hint to pay attention to the format or something like that, so that you know what the reason for it was" (Focusgroup2.2, paragraph 96)
3.2 Elaboration of the response	All text passages that show how improvements in the elaboration of the response can be achieved by incorporating various factors	"that's why I think it would be great to have an explanation, especially if things like elective courses are also included in the calculation or not." (Focusgroup2.2, paragraph 76)
3.3 Logic behind the model	All text passages that demonstrate improvement tips related to the fundamental logic of the model	"that one could either write in that it's the expected credit points achieved by the end of the semester, or that one could indicate the last one" (Focusgroup2.2, paragraph 94)

only a few seconds. In addition, participants appreciated the ease of use and the intuitiveness of the application. However, the sensitivity of the data format was mentioned several times. The calculation may be incorrect even with small deviations from the predefined data format.

The output quality of the study forecast did not fully meet the participant's expectations. Frequently mentioned was the lack of transparency regarding calculating the remaining study time and calculating the credit points still to be completed. A written justification of the result would be considered helpful. It was also mentioned that the display of the remaining credit points needs to consider that the student is still in a current semester. Thus, the number may be incorrect. Moreover, the recommendations' choice of words was felt too harsh. Especially in the case of a deviation from the study plan, a more sensitive wording or, as already mentioned, a detailed justification would be preferred. Finally, it was noted that the result should be supplemented with advice like visiting a counseling center (Table 2).

5 Discussion

In this paper, the development of a study progress forecast was illustrated. During the development process and the focus group several findings could be derived that will be discussed in the following.

The most pressing question considering the existing implementations in the German higher education environment is whether the presented system can monitor study progress adequately. These concerns were also brought forth in the focus group. In literature, numerous systems are mentioned that use multiple indicators to create a study progress prediction. For example, [5] consider over 80 variables in their prediction model, including demographic factors such as gender or family background, as well as aspects related to the first semester or the studying as a whole, such as the type of university or study satisfaction. As a result of this study, it was found that not all factors have an influence on academic progress and that the effect can differ depending on the field of study. It is also worth noting that such models are usually intended for use by the university itself, which uses them to identify students who are at risk of dropping out. The construction of such a system is currently not planned at the University of Ansbach or is complicated by various factors, such as data protection regulations and the presence of different data sources that would need to be migrated. Concerning data protection, there was also the requirement that student data should not be stored in the service. As part of the research project, a lightweight formula was sought that would use a few but significant aspects that the students could easily capture to determine the study's progress.

Another aspect worth mentioning when talking about the study progress forecast is the way students handle information about their academic progress. [17] explored the reaction of new students to feedback on their accumulated achieved academic performance in an experimental study design. Positive feedback led to an increase in academic performance, but no change was recorded with neutral or negative feedback. Therefore, the PASST?! research project highlights the importance of careful wording of the messages [11]. Difficulties in academic progress should not be tabooed, but solutions should be presented [18]. This corresponds with the focus group results. The participating students appreciated the possibility of conducting a study progress prediction, especially its anonymity, but also pointed out the importance of a sensitive formulation of recommendations.

6 Conclusion

The aim of this paper was to present the development and evaluation of a first prototype of a study progress forecast. First, implications for preparing a forecast application from the literature were described. Based on the requirements identified in the literature and from the project itself, a lightweight formula was created that enables the calculation of the remaining study time. The module performing the application is integrated within a conversational interface that can be accessed via the university homepage. All the data that are needed to obtain the remaining study time are entered by the student and will not be stored. To acquire insights about the application, a qualitative focus group was conducted. Aspects that were covered during the discussion were future use, positive aspects and suggestions for improvement. Overall, the study progress forecast was well received by the participants, especially the offer itself as well as the fast output of the service were highlighted. However, the focus group also illustrated some practical implications for the further development of the application, especially regarding the plausibility of the calculation and the wording. Having laid the foundation with the first prototype, further development of the service will be based on the new insights.

Limitations and Potential for Future Work

Some restrictions regarding this study will be mentioned in the following. One of the primary drawbacks of the focus group design is the small size of the participant group. With only 5 participants, it's challenging to consider the group a representative sample of all students. Future focus groups should aim for a more extensive and more diverse group of participants, considering also participants from other universities. Another issue to consider is the presence of a lecturer and a research assistant in the focus group. The original plan was for the group to consist solely of students. However, the inclusion of the lecturer and research assistant became necessary due to difficulties in recruitment. It could be beneficial to also interview lecturers and representatives of the university as users, as they have the potential to use the application for analyzing anonymized data on the study progress of a cohort.

Insights gained from the literature and the focus group allowed for the identification of several possibilities for the further development of the study progress forecast. Although a lightweight formula was intentionally chosen to determine academic progress, there is still the option to expand this progress forecast to additional variables to make it even more accurate. For example, [5] found a strong influence of the score of the higher education admission qualification. This should be evaluated in future studies. It is also of great importance to deal with the issue of the output of the academic progress forecast. So far, the result is only a statement indicating whether the student is in the planned academic schedule or not. The addition of further recommendations appears to be sensible. In addition, the calculation of the remaining study time should be made more transparent by adding an explanation or information about the formula to the answer.

The further development should be investigated in another focus group. Other evaluation formats are also possible. To reach a larger sample, the use of an online-based survey at the entire university is a likely option.

References

1. Heublein, U., Hutzsch, C., Schmelzer, R.: Die Entwicklung der Studienabbruchquoten in Deutschland. DZHW Brief (2022). https://doi.org/10.34878/2022.05.DZHW_BRIEF
2. Heublein, Ulrich und Wolter, Andrä, "Studienabbruch in Deutschland. Definition, Häufigkeit, Ursachen, Maßnahmen (2011) https://doi.org/10.25656/01:8716
3. Neugebauer, M., Heublein, U., Daniel, A.: Studienabbruch in Deutschland: Ausmaß, Ursachen, Folgen, Präventionsmöglichkeiten. Z. Erzieh. **22**(5), 1025–1046 (2019). https://doi.org/10.1007/s11618-019-00904-1
4. S. Anastasio u. a., Studienabbrecher, innen als Zielgruppe der Beratung und Öffentlichkeitsarbeit. DE: wbv Media, 2020. Zugegriffen: 10. Februar,: Online. Verfügbar unter: (2023). https://doi.org/10.3278/6004808w
5. Behr, A., Giese, M., Teguim, H.D., K, Katja Theune,: Early prediction of university dropouts – a random forest approach. Jahrbücher für Nationalökonomie und Statistik **240**(6), 743–789 (2020). https://doi.org/10.1515/jbnst-2019-0006
6. Bundesministerium für Bildung und Forschung, "Bekanntmachung", Bekanntmachung, 26. November 2019. https://www.bmbf.de/bmbf/shareddocs/bekanntmachungen/de/2019/12/2776_bekanntmachung (zugegriffen 5. Januar 2023)
7. Bundesministerium für Bildung und Forschung, "DMPS", DMPS. https://www.wihoforschung.de/wihoforschung/de/bmbf-projektfoerderung/foerderlinien/studienerfolg-und-studienabbruch/studienerfolg-und-studienabbruch-i/dmps/dmps.html (zugegriffen 5. Januar 2023)
8. Schneider, K., Berens, J., Burghoff, J.: Drohende Studienabbrüche durch Frühwarnsysteme erkennen: Welche Informationen sind relevant? Z. Erzieh. **22**(5), 1121–1146 (2019). https://doi.org/10.1007/s11618-019-00912-1
9. Pixner, J., Mocigemba, D., Kraus, M., Krempkow, R.: Sag mir, wo die Studis sind. Wo sind sie geblieben? Outputorientierte Qualitätssicherung auf Studiengangsebene mithilfe der Studienverlaufsanalyse", Das Hochschulwesen, Bd. 57 (2009)
10. Hochschule Mittweida, "Studiengangmonitoring und Wissensmanagement". https://www.hs-mittweida.de/webs/studiengangmonitoring/ (zugegriffen 7. Januar 2023)
11. Schulze-Stocker, Franziska, Blum, Cornelia, Dunkel, Pauline, und Rockstroh, Michael, "PASST?! Partnerschaft Studienerfolg TU Dresden: Das Frühwarnsystem für Studierende an der TU Dresden", S. 292 KB, 300 KB, 5 pages (2021) https://doi.org/10.26204/KLUEDO/6468
12. Bocklisch, T., Faulkner, J., Pawlowski, N., Nichol, A.: Rasa: open source language understanding and dialogue management. arXiv, 15. Dezember 2017. Zugegriffen: 14. Februar 2023. [Online]. Verfügbar unter: http://arxiv.org/abs/1712.05181
13. Schulz, M.: Quick and easy!? Fokusgruppen in der angewandten Sozialwissenschaft. In: Schulz, M., Mack, B., Renn, O. (eds.) Fokusgruppen in der empirischen Sozialwissenschaft, pp. 9–22. VS Verlag für Sozialwissenschaften, Wiesbaden (2012). https://doi.org/10.1007/978-3-531-19397-7_1
14. Fersch, M.-L. , Schacht, S., Woldai, B., Kätzel, C., Henne, S.: Digital Learning Assistants in Higher Education Environments: A Qualitative Focus Group Study. gehalten auf der The Barcelona Conference on Education 2022, Nov. 2022, S, pp. 325–341. https://doi.org/10.22492/issn.2435-9467.2022.28

15. Sek, Y.-W., Law, C.-Y., Liew, T.-H., Hisham, S.B., Lau, S.-H., Pee, A.N.B.C.: E-Assessment as a self-test quiz tool: the setting features and formative use. Proc. - Social Behav. Sci. **65**, 737–742 (2012). https://doi.org/10.1016/j.sbspro.2012.11.192
16. Kuckartz, U.: Qualitative Inhaltsanalyse: Methoden, Praxis, Computerunterstützung, 3., Überarbeitete Auflage. in Grundlagentexte Methoden. Weinheim Basel: Beltz Juventa, 2016
17. Brade, R.: Normatively Framed Relative Performance Feedback – Field Experiment and Replication. American Economic Association. https://doi.org/10.1257/rct.3288
18. Blum, C., Rockstroh, M.: Hinschauen lohnt sich: ein Frühwarnsystem im Interesse der Studierenden und der Universität", Zeitschrift für Beratung und Studium, Bd. 13

AI for Coding Education Meta-analyses: An Open-Science Approach that Combines Human and Machine Intelligence

Vipul Gupta[1(✉)], Brian R. Belland[2], Alexander Billups[1], and Rebecca J. Passonneau[1]

[1] Department of Computer Science and Engineering, College of Engineering, Pennsylvania State University, State College, US
{vkg5164,abb5975,rjp49}@psu.edu

[2] Educational Psychology, Counseling, and Special Education, College of Education, Pennsylvania State University, State College, US
bbelland@psu.edu

Abstract. Meta-analysis provides researchers with a way to assess the efficacy of an educational intervention across multiple independent studies by integrating them into a single statistical analysis, and thereby generalize over a larger, more heterogeneous population. This influences the ability to address goals of diversity, equity and inclusion (DEI), by providing a perspective over different populations of students. However, meta-analysis is extremely costly, mainly due to the need to manually code each of the many articles selected for inclusion, for each relevant variable. To shorten the time to publication, lower the cost, enhance transparency, and enable periodic updates of a given meta-analysis, we propose an open-science approach to meta-analysis coding that provides distinct modules for each variable, and that combines human and automated effort. We illustrate the approach on two variables that represent two types of automated support: pattern matching, versus machine learning. On the latter, we leverage a human-in-the loop approach for a variable that identifies distinct student populations, and is thus important for DEI: we report high accuracy of a neural model, and even higher accuracy of a selective prediction approach that defers to humans when the model output is insufficiently confident.

Keywords: Meta Analyses · Machine Learning · Human-in-the-loop · Natural Language Processing

1 Introduction

Meta-analysis is a method to systematically synthesize the results of multiple research studies that measure the efficacy of a particular treatment, such as a

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023
T. Schlippe et al. (Eds.): AIET 2023, 190, pp. 14–29, 2023.
https://doi.org/10.1007/978-981-99-7947-9_2

medication or an educational intervention. Meta-analysis drives informed decisions and can help practitioners avoid unsatisfying or even catastrophic results from flawed methods. Here we focus on applying meta-analysis to education interventions, where synthesis of hundreds of studies is common. While educational interventions are not as high stakes as medical ones, education is critical to human growth, and should seek the strongest evidence bases. The most time-consuming step involves manual coding of all included studies. Here we present results of a feasibility study to partially automate manual coding for educational meta-analysis.

Research on medical meta-analysis indicates that results are often outdated when published [22,24], and are rarely updated. This makes it hard to understand how research findings evolve as new studies are carried out. Transparency and reproducibility are critical within educational research [14], including within meta-analysis, but is often lacking in meta-analysis reports [29]. We report a feasibility study of automated meta-analysis coding. Our goals are to investigate ways to reduce the time to publication, to support repeating a given meta-analysis as new literature accumulates, and to provide greater transparency.

One of the most time consuming steps of meta-analysis is the manual coding of included articles. Based on our experience, we estimate the person hours for manual coding, including training and inter-rater reliability, to be between 0.25 h and 1 h per coder per variable per article, or between 400 and 1600 person hours per meta-analysis, typically spread over a period of six to ten months.[1] We envision creation of distinct modules for each coding variable in a given meta-analysis, including the input data, output values, and coding software. We refer to such modules as Meta-analysis Coding Replicability Objects (MACROs). Each MACRO would contain the definition and manual coding instructions for a single meta-analysis variable, coded data for included articles, and software to automate the coding of one or more values of the coding variable. Publicly accessible MACROS would be consistent with an open science approach to meta-analysis coding.

Our use case is meta-analysis of educational scaffolding. Scaffolding is an instructional approach based on the zone of proximal development proposed by Russian psychologist Lev Vygotsky (1978) in which students' cognitive skills are iteratively assessed to provide them with enough support, or scaffolding, to help them to the next stage of independent mastery. Through meta-analysis, scaffolding has been found to raise student performance levels substantially with relatively little instructional time, and to be highly effective across STEM disciplines and education levels [4,5]. These meta-analyses identify moderating effects that relate to different student populations, such as urban versus rural, or under-represented minorities, which can therefore inform policies to promote diversity, equity and inclusion. Thousands of new studies on scaffolding in STEM education have been published since the latest article included in [5].

[1] The variability is due to large differences between variables, and between journal article format and clarity.

We present results for two types of autocoders, each for a different variable drawn from [5] that are likely to be relevant for many meta-analyses within education.[2] Pattern-based-autocoding can be applied to variables where coding values are objective properties that are explicitly expressed in one or more keyword phrases in the article, and which can be detected through pattern matching. Education level (henceforth `ed-level`) falls into this category: coding values can be identified in phrases such as "seventh grade students" or "intermediate level under-graduate students." Our pattern-based autocoder for `ed-level` achieved 85% accuracy. Natural language processing is required for variables whose values must be inferred from indirect textual evidence (learning-based-autocoding), such as education population (henceforth `ed-pop`). For example, the choice of under-represented when most participants are female depends on the subject matter, such as physics versus biology. Our `ed-pop` classifier had an accuracy of 93.75%. As part of a human-in-the-loop system accuracy was 96.53%.

2 Related Work

Automated support for extracting coding categories for meta-analysis has been explored within the medical literature. Effect sizes of treatment can often be extracted from the abstracts, which in the medical literature commonly have an explicit, known structure [2,8,22,25,39]. In contrast, only 11% of the top 150 education journals according to impact factor require structured abstracts [10]. Related tools used in the medical meta-analysis, such as EXACT [30] or a tool using crowd-sourced annotated abstracts [26] would not apply to educational literature for similar reasons. However, these kinds of tools have led to sizable time savings: medical researchers who used automated tools that leverage structured abstracts completed a publishable meta-analysis report within 2 weeks [9]. It will be more challenging to automate aspects of education meta-analysis, but our study shows a path forward.

To our knowledge, no prior work in educational research applied NLP to meta-analysis variable coding. Thus, there is no standard problem formulation and no publicly available datasets. Open-domain question answering (OpenQA) is a related task with a large literature. Given a question and a collection of documents, the goal is to find an answer to the question. The problem formulation consists of retrieval of relevant documents from the collection, selection, and encoding of one or more sentences containing the answer, then generation (decoding) of one or more novel sentences to answer the question. The first step is irrelevant for meta-analysis because the journal articles have already been identified. Instead of a final decoding step, an output classification layer can assign encoded input to a coding variable value. Highly cited approaches include [19,32]. The state-of-the-art Fusion-in-Decoder method from [17] has a simple architecture: all retrieved passages for a query are separately encoded, then a decoder performs attention over the concatenation of the encoded passages. This informs the model we use in our study.

[2] Data and code will be made available upon acceptance.

Instead of fully automated coding, we propose to automatically code values where we can achieve high performance, and to combine human effort with autocoders where humans code the residue of uncoded articles. Our study provides an example variable where this might apply. In addition, there are multiple ways to design human-machine collaboration for semi-automated coding, including human-in-the-loop, where humans provide support to algorithmic decision-making, and algorithm-in-the-loop, which favors human decision-making. Selective prediction, where a classifier defers to a human when the system decision (i.e., class value) has low confidence, has been found to be an effective human-in-the-loop approach [7,15,43]. Although rarely used in NLP, Xin et al. (2021a) presented a regularizer for pre-trained transformers like BERT [11] that improves confidence estimation, for the decision of when to defer to human decision makers.

A	Coding variable	Name and identifier
B	Brief specification	List of coding values and short description
C	Full specification	Criteria for each value with illustrative examples
D	One or more 5-10 min. training videos	Tutorial demonstration of challenging examples
E	Articles	Table of article citations with indicators of which were manually and/or automatically coded, and associated meta-analyses
F	Manual coding results	Human coding of articles (E), *with supporting sentences*
G	Quality metrics on manual coding	Inter-rater agreement and person hours
H	Text extraction instructions and code	Python library to convert article pdfs to plain text; to post-process extracted text into a simple XML format
I	Software for automatic coding	Link to python code repository to autocode post-processed text (F), including specific train/test split for machine-learned autocoders (F)
J	Autocoding results	Output of the module code (G) on relevant articles (E)
K	Quality metrics on autocoding	Autocoder (H) accuracy, agreement with humans, and efficiency
L	Meta-analyses	A list of published or public meta-analyses utilizing this coding variable, and a table for each indicating rows of E, F and J used.

Fig. 1. Elements of a Meta-Analysis Coding Replicability Object (MACRO)

3 Background and Goal

3.1 Meta-Analysis

The major steps for any meta-analysis are largely consistent: (1) Front-end work: Operationalizing constructs and creating search terms and coding scheme, (2) Searching for literature, (3) Applying inclusion/exclusion criteria, (4) Coding included articles, (5) Running meta-analytic analyses, and (6) Writing the meta-analysis report. This project will use natural language processing techniques to fully or partially automate coding.

One of the greatest challenges for meta-analysis is the high cost: it is labor-intensive and thus both expensive and time-consuming. In [22,24,40], it was found that a systematic review took on average 67.3 weeks (range: 6–186 weeks) for an author team of 5 persons. The startup time of a meta-analysis before locating and screening articles is estimated to be 721 h [1]. A recent study estimated that each of the 10 institutions receiving the highest amount of funding from the National Institutes of Health (NIH) spent on average $18,660,305 every year producing meta-analyses [24]. The 10 institutions together spend about $186,603,050 each year producing meta-analyses. Presumably, most of this money is from federal and other grants. The 10 largest pharmaceutical companies each spent on average $16,761,235 every year on meta-analyses [24], which presumably contributes to medication cost.

Deprecation of results occurs in part as a byproduct of the high cost in time. Between the time a literature search is completed and the time of publication, new, relevant results are often published, leaving a newly published meta-analysis already out of date [6,27]. In medicine, the median survival time for meta-analyses before a need for updating was 5.5 years [36]. Through automated support for manual coding, our approach has the potential to reduce the cost and time-to-completion, and support periodic updates.

Another challenge to meta-analysis is coding bias, which can arise when coders have different interpretations of coding schemes or approaches to coding [18]. This is less of a problem when coding low-inference items like `ed-pop`. It is more of a problem when coding high-inference items, where the coder must actively interpret the meaning of the text. One can assess coding using chance-adjusted measures of inter-rater reliability like Krippendorff's alpha [21]. Our approach to automated support includes the goal to achieve at least the same accuracy as human coders through human-in-the-loop methods. We do not address publication bias. Our study includes an example of a low-inference item and a high-inference item.

3.2 Scaffolding

Scaffolding is widely in STEM education because it helps students learn to solve problems and engage in argumentation while engaging in the types of tasks in which STEM professionals engage, like debugging programs and finding engineering solutions to problems [13,16,20]. It does this through a combination of cognitive and motivational support that elicits interest, control frustration, and structures the problem-solving process [3,34,41]. The term was originally applied to education in the context of teachers working one-on-one with young children to help them build pyramids using wooden blocks [42]. In this process, teachers enlist and maintain student interest, simplify the problem and highlight important problem elements, model effective problem-solving strategies, and adjust (fade and/or add) scaffolding as student abilities change. Critical to the definition is that scaffolding (a) is provided to learners as they engage in an ill-structured problem or task, and (b) supports current performance and leads to skill gain.

Existing theory and empirical studies have suggested that fading scaffolding leads to better learning outcomes than not fading [23,31]. Still, the full range of scaffolding change types (i.e., fading, adding, fading/adding, and none) and change bases (i.e., performance-based, self-selected, fixed, and none) had not been compared empirically until the scaffolding meta-analysis published by [5], which found no difference in cognitive outcomes based on scaffolding change type. As noted above, there have been thousands of new studies of scaffolding in STEM since that study.

3.3 MACROs

The main vehicle for our open-science approach to meta-analysis coding variables consists of Meta-Analysis Coding Reproducibility Objects (MACROs). Just as meta-analysis works best when authors of original studies are transparent and open in the reporting of their research, so too can the meta-analysis community move forward the best when meta-analysts are also open and transparent in reporting their work [28]. The term MACRO is intentionally reminiscent of the use of the term macro in computer programming for a subroutine that supports the re-use of a common set of operations. We envision a public repository of MACROs, where each defines a coding variable, provides resources to apply it manually or with automated support, and stores the coded output of one or more sets of articles, along with quality metrics. These elements of a MACRO contribute to the transparency of the manual coding, documentation of evolving best practices for coding, coder training, re-use of the same MACRO across different meta-analyses, and iterative evolution of a meta-analysis as new literature accumulates. Therefore MACROs would promote open science more fully than the release of the coding dataset, or software for automated coding.

Figure 1 illustrates the structure of a MACRO, corresponding to a single module for a given meta-analysis or set of related meta-analyses, such as computer-based scaffolding in STEM education [4,5]. In this paper, we focus on the design and performance of two MACROS and leave for future work the question of how best to host a set of MACROS that supports mix-and-match utilization. Items A-G in Fig. 1 provide transparency and support training for manual coding. Item D is new, whereas Items B, C, and F are typically created in a meta-analysis, but not publicly shared. For Item G, we include inter-rater agreement, which is usually included in a published meta-analysis, but we also include person-hours of effort, which is typically omitted due in part to length limitations. Items H-L support the automation of one or more coding values. Items E and F can go beyond a conventional meta-analysis to include articles and manual coding that are for articles excluded from the meta-analysis, but that are similar enough to include articles to support training and testing of autocoders, especially those that depend on machine learning. One of the two MACROs we discuss here falls into this category. Item E, the table of coded articles, will include all articles that support at least one meta-analysis along with articles that were not used in a meta-analysis, but were coded to provide data to develop autocoders. Item F, the manual coding results, includes the coding values, and optionally, one or more sentences from each article that support the coding decision. This adds extra effort to the manual coding but can provide a payoff in facilitating machine learning approaches.

The item I, autocoder software, is the most novel component, so its purpose merits a clear statement. The function of autocoding is to significantly reduce human effort through rule-based or machine-learned software, rather than to replace human effort. We envision many ways that an autocoder can reduce human effort. A highly accurate autocoder could potentially replace human effort for all values of certain variables. However, an autocoder might have sufficiently high accuracy for only a subset of variable values. The autocoding quality metrics (K) allow users to decide whether an autocoder is sufficiently accurate to use as a standalone tool, or in combination with human effort, e.g., as part of a human-in-the-loop system [7,15,43]. Below, we present the results of an experiment to develop a machine-learned autocoder for one of the two variables discussed here, and combined with an offline test of the feasibility of human-in-the-loop. The value in question accounts for approximately 80% of the cases, thus human effort is required for the human-in-the-loop, and for the 20% of remaining cases.

Table 1. Example regular expressions showing strings each would match, and the candidate `ed-level` value.

RegEx	Matching strings	Ed-level
(seven|7)th grade(|ers)	seventh grade, seventh graders, 7th grade, 7th graders	Middle sch.
grade[s]* (seven|7 |6 and 8)	grade seven, grade 7, grades 6 and 7	Middle sch

4 Two Example Autocoders

To investigate the feasibility of autocoders for MACROs, we selected two variables from a previously published meta-analysis on scaffolding in STEM education [5]. The two variables were selected based on the following criteria: both variables are likely to be relevant for many education meta-analyses; they represent two types of variables, low-inference, and high-inference (see above), corresponding to two types of autocoders; together they provide a way to assess to what degree the studies investigate diverse groups of students, and thus how well scaffolding methods work for different subpopulations. The initial data for both autocoders comes from the coding workbook (WB) for the [5] meta-analysis. The WB documents the manual coding results, provided by the first author of that study, who is the second author of this paper. Here we briefly describe the WB data, the two types of autocoder, and the preprocessing of articles that generates the input for both autocoders. The two subsections that follow describe the autocoders in turn.

The information we extracted from the manual coding workbook (WB) consisted of the citation information for each of the 144 publications included in the [5] meta-analysis, and the consensus coding for the ten variables of interest. We treat the consensus coding values as ground truth for developing the two autocoders.