David Buschhüter[1]
Marisa Pfläging[1]
Andreas Borowski[1]

[1]Universität Potsdam

Uncovering Statistical Relationships Using Structural Topic Models in Science Education

New ways machine learning can contribute to science education research

Studies have shown that using unsupervised machine learning techniques opens promising ways to contribute to science education theories directly. Sherin (2013) identified new categories regarding student conceptions on physics explanations of the seasons. Similarly, Rosenberg and Krist (2021) used unsupervised methods to get to a more nuanced category system regarding students' ideas of the generality of model-based explanations. These studies show the intriguing potential of methods extracting categories a human might not have found. However, our theories also consist of relations between such entities. Thus, it would be interesting to extract categories related to other variables of the text automatically. This would represent a prototypical way of contributing to the literature and would make finding categories of interest more probable.

In principle we can accomplish the task above using so-called structural topic models (STMs) (Roberts, Stewart & Tingley, 2019). These models allow us to extract the topics from a corpus of documents given metadata (covariates) for the documents. Various studies employed STMs outside the field of science education research (e.g. Roberts et al., 2014). For the educational context, Reich, Tingley, Leder-Luis, Roberts and Stewart (2015) used this method in order to investigate massive open online courses. It also has been used to identify trends in science education research over time (Mi, Lu & Bi, 2020). So far, there are, to our knowledge, no studies applying this approach to science education research questions. Therefore, we are presenting two cases in which we applied STMs to answer the question: To what extent can STMs help us extend our theories by supporting the search for new categories and relationships? We provide an exemplary answer to this question by presenting two use cases with varying data set sizes and German language text data.

Use case I

The data consists of 102 booklets of the VNOS-C (Views of Nature of Science (form C)) questionnaire filled out by in-service teachers (Lederman, Abd-El-Khalick, Bell, & Schwartz, 2002, translated into German by Hofheinz, 2008, and Pfläging, in preparation) before and after a teacher development program on NOS (based on Pfläging, Richter, & Borowski, 2020). The

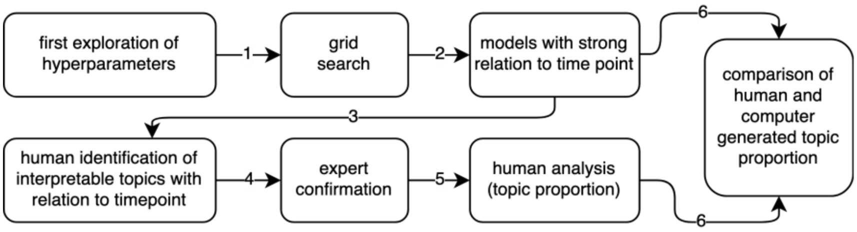
first exploration of hyperparameters
1
grid search
2
models with strong relation to time point
6
comparison of human and computer generated topic proportion
3
human identification of interpretable topics with relation to timepoint
4
expert confirmation
5
human analysis (topic proportion)
6

Figure 1: Analysis Workflow

goal of the analysis was to explore if we can extract topics from the response texts which show a statistical relationship to the variable *time point* (pre or post) in order to identify meaningful learning processes. We used several standard pre-processing steps (e.g., stemming, stop word removal) and employed the r package stm (Roberts et al., 2019) to fit the topic models. A simplified analysis workflow can be seen in Figure 1. We performed a grid search on a set of hyperparameters (e.g. number of topics 2-10) with *time point* as covariate and analyzed the VNOS-items separately. Using R Markdown (Allaire et al., 2021), we created reports for all topic models with a significant relationship between any of the topic-prevalences and *time point.* For models containing high regression coefficients ($|B| > 0.15$) one rater judged whether the topic model (but mostly the topic(s) with a relation to the *time point*) could be interpreted using the most prevalent words for every topic and the responses (documents) with the highest and lowest proportions. We confirmed that the author of the development program agreed that it was plausible that the teachers learned or unlearned the topic-related content during the teacher training. Next, the segments were coded by the rater and topic proportions were calculated. Related to question 2 ("What is an experiment?"), we could confirm the significant difference via human analysis. The topic was previously identified as addressing the role of variables (most prevalent words: experiment, investigation, variables, impact, method, posed the question, variable). After minor post-hoc corrections to the human codes the change in human topic proportion from pre (*Mdn* = 0) to post (*Mdn* = 0) showed to be significant in a sign test ($p = 0.039$, *Mdn* = 0, success ratio 10/12) and was correlated to the machine-based topic proportion ($r = 0.372$, $p < 0.001$, $t(82) = 3.624$, $N = 84$). A higher correlation ($r = 0.410$) could be reached by including codes that express the idea of variable control but do not use the proper wording (e.g., "In some cases, you have to keep individual things that have an influence constant or change them systematically"). However, then, we did not observe the statistical relation anymore ($p = 0.359$, *Mdn* = 0, success ratio = 12/19).

Use case II

The data of this study consists of 584 answers of physics students on a content knowledge assessment (Enkrott, 2021). The unit of analysis was a constructed response item that prompted students to provide their reasoning for a single-select problem. Here, the students were asked "Which of the following equations is most fundamentally different from the other three from a physics perspective?". The options were four equations ($E_{pot}+E_{kin} = E_{pot}'+E_{kin}'$ and similar equations for mass, angular momentum and force). A supposable solution was to identify *force* because it is not conserved. However, there are other acceptable answers to this question. The goal of the topic modeling approach was to identify different levels of understanding in the student answers. Therefore, we used the *choice* and the total test-*score* as covariates *(prevalence~choice+score)*. Using a reporting system, we identified a model with five topics to be most interpretable. However, the overall topic model was still difficult to interpret, as the five topics needed to be understood for every choice (e.g., energy, mass). In particular, it was difficult to understand why the topic of "conservation of energy, momentum, force" was positively related to the score (Figure 2) because claiming that energy is conserved and therefore choosing the option "energy" is not correct (angular momentum is also conserved and the equation displays conservation of mechanical energy). Therefore, we fit separate sets of models (using *prevalence~score)* for the different subsets of choices (1-4). Here we identified interpretable topic models, and used the topic models as a coding-support: We used the maximum topic proportion to identify one dominant topic for every response, then ordered the responses by dominant topic. Like this, we achieved a prior order. Based on

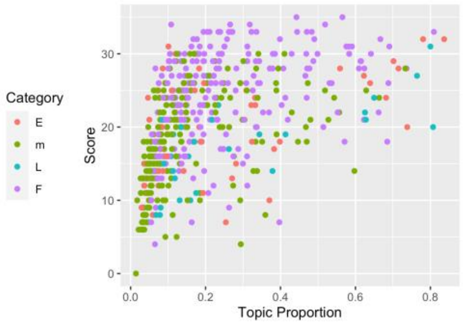

Figure 2: Scatterplot of topic proportions and test score. The category represents the student's choice on the single select item (E = energy, m = mass, L = angular momentum, F = force). The most common words were: conservation of energy, applies, always, conservation of angular momentum, system, conserved, conservation of forces, stays, fundamental, applies (English words mapped to German stems)

this order the human rater could more efficiently form new categories (e.g. "Energy is always conserved"). We ranked and assigned the codes to five different levels of quality. On average all but one level (level 0) appeared in the order anticipated. For students choosing the option "energy", this issue mostly consisted of 16 answers with a code like "Energy is always conserved", which on average had a higher test score than anticipated and which on average also had a higher proportion of the topic "conservation of energy, momentum, force" in our initial topic model, which we discarded. Based on this observation and the initial topic model, we would hypothesize that having an expert perspective of the understanding which properties of entities are fundamental might be more critical in the understanding of physics than giving a more correct but superficial answer (like e.g. "E_{kin} and E_{pot} have different equations but the others are the same", which we initially considered a better way of responding to the question).

Discussion

The first example provides evidence to claim that STMs could help identify learning processes from texts in science education in an exploratory manner. Here the hypothesis extracted is that teachers learn about the role of variables in the experiment during the development program. This aspect is typically not evaluated by the VNOS-C scoring rubric (Lederman et al., 2002) while control of variables is of interest to science education research (Schwichow, Croker, Zimmerman, Höffler & Härtig, 2016).

The second example illustrated how STMs could also help to contribute more directly to theory. Here they supported the derivation of the hypothesis that having a hierarchy of fundamentality in physics seems to be a predictor of physics performance. This hypothesis could be further investigated and might help specify and model the type of meta-knowledge some researchers have called "deeper school knowledge" (Enkrott, 2021). Even if the results provide evidence that STMs help us find relations that are interpretable and interesting to researchers, we believe that performing additional analysis with a second human rater is necessary here. In this sense, our results are preliminary but remain promising. Additionally, we should be aware that we could be "reading tea leaves" (Chang, Boyd-Graber, Gerrish, Wang & Blei, 2009) or find statistical relations by chance (Janczyk & Pfister, 2013). We believe that the procedures employed here do only lead to hypotheses which need further confirmation.

References

Allaire, J.J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A. et al. (2021). rmarkdown: Dynamic documents for R. https://github.com/rstudio/rmarkdown

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. & Blei, D. (2009). Reading tea leaves: How humans interpret topic models. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams & A. Culotta (Hrsg.), *Advances in Neural Information Processing Systems* (22).

Enkrott, P. (2021). Entwicklung des fachlichen Wissens angehender Physiklehrkräfte. University of Potsdam.

Janczyk, M. & Pfister, R. (2013). Inferenzstatistik verstehen. Berlin Heidelberg: Springer.

Lederman, N.G., Abd-El-Khalick, F., Bell, R.L. & Schwartz, R.S. (2002). Views of nature of science questionnaire: Toward valid and meaningful assessment of learners' conceptions of nature of science. Journal of Research in Science Teaching, 39 (6), 497–521. doi:10.1002/tea.10034

Mi, S., Lu, S. & Bi, H. (2020). Trends and foundations in research on students' conceptual understanding in science education: A method based on the structural topic model. Journal of Baltic Science Education, 19 (4), 551–568. doi:10.33225/jbse/20.19.551

Pfläging, M., Richter, D. & Borowski, A. (2020). Entwicklung einer Fortbildung zur Veränderung des Wissenschaftsverständnisses. In S. Habig (Hrsg.), Jahrestagung der Gesellschaft für Didaktik der Chemie und Physik in Wien 2019. Naturwissenschaftliche Kompetenz in der Gesellschaft von morgen 40 (S. 1059–1062). Universität Duisburg-Essen.

Reich, J., Tingley, D., Leder-Luis, J., Roberts, M.E. & Stewart, B. (2015). Computer-assisted reading and discovery for student generated text in massive open online courses. Journal of Learning Analytics, 2 (1), 156–184. doi:10.18608/jla.2015.21.8

Roberts, M.E., Stewart, B.M. & Tingley, D. (2019). Stm: An R package for structural topic models. Journal of Statistical Software, 91 (2). doi:10.18637/jss.v091.i02

Roberts, M.E., Stewart, B.M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S.K. et al. (2014). Structural topic models for open-ended survey responses. American Journal of Political Science, 58 (4), 1064–1082. doi:10.1111/ajps.12103

Rosenberg, J.M. & Krist, C. (2021). Combining machine learning and qualitative methods to elaborate students' ideas about the generality of their model-based explanations. Journal of Science Education and Technology, 30 (2), 255–267. doi:10.1007/s10956-020-09862-4

Schwichow, M., Croker, S., Zimmerman, C., Höffler, T. & Härtig, H. (2016). Teaching the control-of-variables strategy: A meta-analysis. Developmental Review, 39, 37–63. doi:10.1016/j.dr.2015.12.001

Sherin, B. (2013). A computational study of commonsense science: An exploration in the automated analysis of clinical interview data. Journal of the Learning Sciences, 22 (4), 600–638. doi:10.1080/10508406.2013.836654