Expanding physics education understanding through large-scale literature review using unsupervised natural language processing

Martina CARAMASCHI (1), Tor Ole B. ODDEN (2), Olivia LEVRINI (1)

(1) *Department of Physics and Astronomy "A. Righi", University of Bologna, 40126 Bologna, Italy*
(2) *Center for Computing in Science Education, University of Oslo, 0316 Oslo, Norway*

Abstract. We have thematically analysed papers from "The Physics Teacher" journal from its inception in 1963 to 2020, to understand what themes were prevalent in the history of physics teaching. The methodology combined an unsupervised machine learning (ML) method called Latent Dirichlet Allocation (LDA) into a qualitative analysis. Specifically, LDA allowed us to identify patterns of words that represent "topics", while researchers derived analytical process and interpretation of results. Our analysis found 13 topics displayed over time, grouped into content-focused topics, pedagogical, laboratory and data analysis-focused topics, and learning-theory topics, suggesting a shift from practices to considering relevant learning theories.

Introduction, theoretical framework, and research question

Studying the evolution of physics teaching practices and interests over time is a complex challenge. Systematic literature reviews on physics teaching can contribute, but face limitations like the effectiveness of qualitative methods for analysing large text datasets in a reasonable time and the ability to visualise and quantify trends. In recent years, the use of ML and natural language processing (NLP) methods to support qualitative analyses of text-based data has taken hold. In the Physics Education Research (PER) field, two works used ML and NLP techniques to thematically analyse conference proceedings and journal articles [1][2], enabling us to find topics trends, visualising and quantifying them. Our work aimed to analyse a wide corpus of papers published from 1963 to 2020 in *The Physics Teacher*, a peer-reviewed academic journal, to explore what topics have been of interest during its history. By including NLP and ML methods in the qualitative analysis conducted by PER experts, we aimed to quantify and visualise topic trends. Following Grimmer and Stewart's second principle for text analysis of using quantitative methods to enhance researchers' capabilities. Our research question was: "What topics are present in the history of *The Physics Teacher* journal papers and to what extent are they prevalent each year?".

Method, results and conclusion

The methodological process has firstly involved a pre-processing phase, where papers are downloaded, and transformed from PDF documents to the corpus of words that compose them, exploiting NLP techniques. Secondly, we performed the topic discovery using LDA. Assuming that each paper contains a mixture of topics, and each topic is a distribution over words, the LDA model infers the topics that best explain the content of the documents [4]. Lastly, the topics' lists of words have been interpreted, based on the titles and texts of the articles that primarily contain each topic, and the literature on physics teaching [5]. All these phases were performed on Jupyter Notebooks written in Python language. The optimal number of topics found was 15. Two were removed since were related to journal business (e.g. awards, announcements). The analysis allowed us to quantify the average prevalence of each topic in the journal year by year. For example, Topics 10, 11, and 12 about pedagogy and laboratory activities in Physics seem to significantly increase their prevalence between 1980 and 2000, when it stabilises (Fig. 1). Observing topics' meaning, trends, and reference literature, we grouped topics into three categories. The first one, content-focused topics, includes those describing physics knowledge and contents (e.g. astronomy, thermodynamics, optics). An

example is Topic 1 on "Particle Physics," with topic representative words like "energy," "particle," "electron," and "atom." An article that is 92% represented by this topic is entitled “A Centennial of Protons”. In general, articles that are mainly described by content-focused topics inform about new physics discoveries, applications for physics teaching, and curiosities. This group of topics have been consistently present (or slightly decreasing) over time. Then, there are the pedagogy, laboratory, and data analysis-focused topics, which emerged around 1985, and the learning-theory topics, which started to trend on late 1980s, eventually becoming the most prevalent topics in the journal (Fig. 2). The articles that mainly belong to pedagogy, laboratory, and data analysis-focused topics are characterised for being more explicitly driven by the intent of describing pedagogical activities, the use of laboratory and experiments for teaching. Instead, in the articles that belong to learning-theory topics there is an explicit reference to learning theories for Physics education.
These results support the idea that the literature has balanced content understanding, pedagogy (i.e., teaching strategies) and learning perspectives, but students' learning has become more valued over time. Besides, this work underlines NLP usefulness in PER for large-scale, inductive, and quantitative literature reviews.

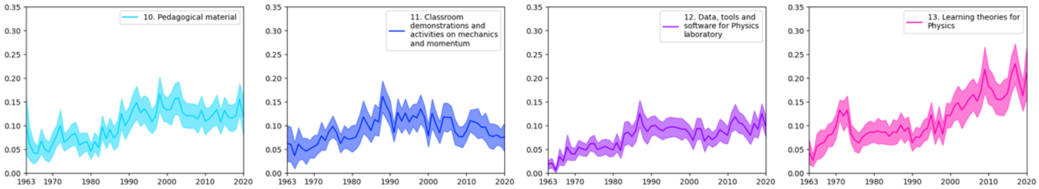

Fig. 1. Example of four different topics: average prevalence in the articles year by year.

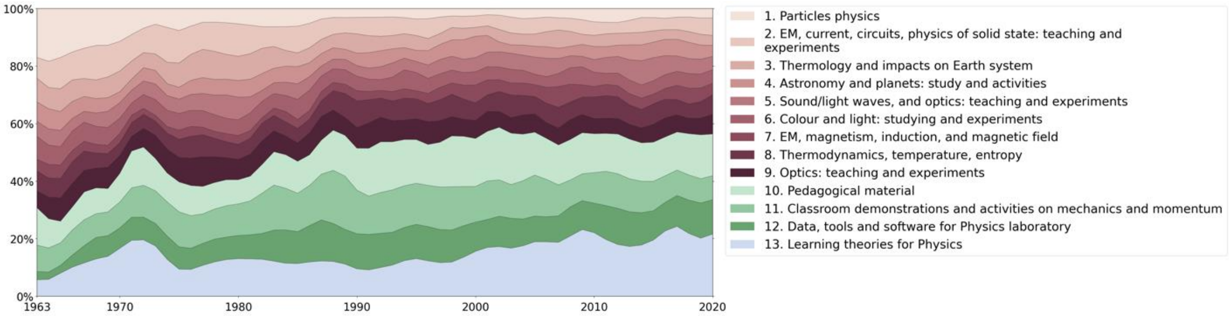

Fig. 2. Stacked area plot of all topics, grouped according to their thematic clusters: physics content topics (red, top), pedagogical/laboratory topics (green, middle), learning theories/PER topic (blue, bottom).

References

[1] T. O. B. Odden, A. Marin and M. D. Caballero. Thematic analysis of 18 years of physics education research conference proceedings using natural language processing, *Phys. Rev. Phys. Educ. Res.* **16** (2020) 010142, 1-25 doi: 10.1103/PhysRevPhysEducRes.16.010142

[2] T. O. B. Odden, A. Marin and J. L. Rudolph. How has Science Education changed over the last 100 years? An analysis using natural language processing, *Sci. Ed.* **105** (2021) 653–680. doi: 10.1002/sce.21623

[3] J. Grimmer and B. M. Stewart. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts, *Polit. Anal.* **21** (2013) 267-97. doi: 10.1093/pan/mps028

[4] D. M. Blei, A. Y. Ng and M. I. Jordan. Latent Dirichlet allocation, *J. Mach. Learn. Res.* **3** (2003) 993–1022. http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

[5] D. E. Meltzer and V. K. Otero, A biref history of physics education in the United States, *Am. J. Phys.* **83** (2015) 447. doi: 10.1119/1.4902397