Developing a natural language processing approach for analyzing student ideas in calculus-based introductory physics

Jon M. Geiger and Lisa M. Goodhew
Department of Physics, Seattle Pacific University, 3307 3rd Ave W, Seattle WA, 98119

Tor Ole B. Odden
Department of Physics, Center for Computing in Science Education, University of Oslo, 0316 Oslo, Norway

Research characterizing common student ideas about particular physics topics has significantly impacted university-level physics teaching by providing knowledge that supports instructors to target their instruction and by informing curriculum development. In this work, we utilize a Natural Language Processing algorithm (Latent Dirichlet Allocation, or LDA) to identify distinct student ideas in a set of written responses to a conceptual physics question, with the goal of significantly expediting the process of characterizing student ideas. We preliminarily test the LDA approach by applying the algorithm to a collection of introductory physics student responses to a conceptual question about circuits, specifically attending to whether it is useful for characterizing instructionally-relevant student ideas. We find that for a large enough collection of student responses ($N \approx 500$), LDA can be useful for characterizing the ideas students used to answer conceptual physics questions. We discuss some considerations that researchers may take into account as they interpret the results of the LDA algorithm for characterizing student's physics ideas.

2022 PERC Proceedings edited by Frank, Jones, and Ryan; Peer-reviewed, doi.org/10.1119/perc.2022.pr.Geiger

I. INTRODUCTION

Over the last few decades, Artificial Intelligence (AI) has been increasingly useful in day-to-day life. From recommendation algorithms on popular streaming streaming services and e-commerce platforms [1] to the programming for self-driving cars [2], applications of AI are vast and ever-growing. Natural Language Processing (NLP) is a branch of AI which has been defined as: "a theoretically motivated range of computational techniques for analyzing and representing naturally-occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications" [3]. With advancements in computational power, NLP has been utilized to analyze enormous amounts of information in a short amount of time.

One application of NLP is known as Topic Modeling [4], which is used to extract themes or "topics" from large bodies of text. Latent Dirichlet Allocation (LDA) is a popular topic modeling algorithm which takes in a set of documents (called a "corpus" in the language of LDA) and produces clusters of words ("topics") that are commonly used together within those documents [4]. From this output, researchers can then ascribe meaning to each of the topics produced by the algorithm. LDA has been applied in fields such as software and banking [5, 6], as well as in education research. In the past few years, there have been several studies exploring the utility of NLP in physics education [7, 8]. LDA has been utilized to characterize topics in Physics Education Research (PER) articles over the last several decades as a way to understand the scope and breadth of the field [9, 10].

A research focus in PER is the investigation of common, topic-specific knowledge that students bring to the classroom. This kind of research has important impacts on physics instruction, particularly at the university level: it informs the development of research-based instructional materials (e.g., Tutorials in Introductory Physics, Maryland Open-Source Tutorials [11, 12]) and it contributes instructors' knowledge of student ideas, which is an important part of the knowledge that instructors use to teach [13]. Research identifying students' common physics ideas has investigated both students' common, incorrect ideas [14–16] and, less extensively, students' common, potentially-fruitful ideas [17–21]. The work of appropriately characterizing resources is very time-intensive; two independent coders may spend upwards of 10 hours each to create and assign descriptive codes to a set of 500 written responses, and often data sets are much larger—this may limit the extent and impact of this kind of research.

The time-intensity of characterizing student ideas motivates the work presented here. This study presents proof-of-concept that LDA may be a useful tool to identify common, instructionally-relevant student ideas. We illustrate a method for automating part of the process of characterizing student ideas by applying LDA to a corpus of student responses to a particular conceptual physics question. We inspect the words contained in a topic determined by the LDA model, then examine documents that best represent that topic to characterize distinct student ideas used in the corpus. This paper builds on previous LDA research by using this approach to analyze students' written responses to physics homework questions, which often include a mixture of technical and informal language.

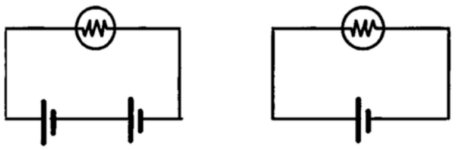
The questions that guide our research are: *To what extent can LDA be used to characterize patterns in student thinking? Can we produce a useful, time-saving method for researchers which yields "instructionally-useful" student ideas?*

II. METHODS

A. Student Task

For this study, we analyzed a set of $N = 483$ written student responses to the conceptual circuits question shown in Figure 1. This question was adapted from [22] as part of a larger study investigating common, instructionally-relevant ideas (e.g., [20, 21, 23]). Responses to this question were an appropriate data set for our this study because: (a) this is a single-part question, (b) responses were text only (not numbers, equations, or diagrams), and (c) a large number of responses to this question had been collected electronically in CSV file format. This question was administered to students as part of a required, online homework assignment in an introductory physics course at a large research university in the United States. Participating students consented to share written work completed for their physics course with researchers, but students were not told which questions were administered for research purposes. Therefore, we assume students framed this question in the same way they framed any other required online homework problem.

FIG. 1. Student task: Explain why a light bulb connected to two batteries is brighter than a light bulb connected to one battery.



> *Comparing the brightness of the bulbs in Circuits 1 and 2, we observe that the bulb in Circuit 1 is brighter. Using Ohm's Law, $V = IR$, we know that current and voltage are directly proportional when resistance is the same. Why do you think more voltage leads to more current? What mental models are you using to make sense of this?*

We used Gensim, a Python package that includes an LDA algorithm, as our primary tool to analyze students' responses to this question [24].

B. Data Cleaning

The following steps were implemented in Python using the "pandas," "re," and "nltk" libraries to prepare the data for LDA modeling [25, 26]:

1. Removed punctuation (quotes, commas, periods, and parentheses).
2. Removed stopwords (commonly used words such as "a," "the," and "is").
3. *Lemmatized* words down to their roots ("increased" and "increases" would become "increase").
4. Created *bigrams*, (pairs of words such as "potential_difference" or "ohm_law"). This is necessary in order to distinguish between concepts such as "potential difference" and "potential energy," which have distinct meanings in physics, even though they both contain the word "potential."
5. Filtered out the most and least common words (based on user-defined thresholds). This process is explained below.
6. Created a *bag-of-words* for the LDA function [27].

The most and least common words were filtered out (step 5) based on criteria chosen by the researchers. Filtering out the most common words involves choosing a threshold percentage of all documents in which a certain word occurs (the "no above" threshold). In a circuits question, for example, the word "current" may appear in 70% of all responses; a researcher may choose to eliminate all words which appear in more than 50% of all documents to remove "current" and other very common words. The LDA algorithm tends to focus on the most common words and will consequently include them in multiple topics. Removing very common words is a standard step in LDA analysis, as it allows us to discover distinct patterns that appear without those most common words [10]. For our data set, many of the most common words appeared in the problem statement itself, and thus were unhelpful in characterizing unique student ideas.

Filtering out the least common words involves choosing a minimum number of documents in which a word can occur (the "no below" threshold). This step is important because it removes noise that would otherwise increase computation time without yielding additional insight. For example, words that appear in one or two documents may include fanciful words used by only one student, or misspellings of words ("increaes" rather than "increase"). More common misspellings or typos (such as "becuase") tend to appear in more than just one document, and may be preserved with too low a threshold. Given the comparatively small corpus size we used ($N < 500$), we chose to exclude words that are only found in two or three documents. For the analysis presented here, we chose the following model parameters:

- "No above" = 50%. Words like "current" and "voltage" appeared in more than 50% of all responses, and would not be useful in characterizing student ideas.
- "No below" = 3. Setting this threshold any higher would have excluded important words potentially significant for student ideas, such as "gravitational."

C. LDA Modeling

In brief, the LDA algorithm works by taking in a corpus (or collection) of documents, noticing groups of words that commonly occur within certain documents, and picking out those groups of words, labeling them as "topics" [4]. The words within a topic are weighted according to their prevalence in a topic, relating to how often those words co-occur with other words in a given topic. More technically, LDA iteratively "learns" topics by creating and adjusting a probabilistic model for how words are distributed in topics, as well as how topics are distributed among documents.

The LDA modeling process relies on a few key assumptions. LDA assumes that the order in which words occur in a document doesn't matter, nor does the part of speech, etc. LDA also assumes that each document is composed of a weighted mixture of all the topics. During the modeling process, each document is assigned a set of weights indicating how relevant to each topic that document appears to be. That is, for a three-topic model, a given document could be comprised of 70% Topic 1, 20% Topic 2, and 10% Topic 3. This is important for interpreting the results of our model because the "distinctness" of the topics is affected by input parameters which can be specified by the researcher.

The mathematics of LDA is founded upon the Dirichlet distribution, which can be thought of as a multivariate generalization of the beta distribution [28]. With k topics chosen, the Dirichlet distribution is formulated using a $(k-1)$ simplex, existing in k dimensions. A 3-topic model yields a 2-simplex, which is an equilateral triangle projected onto three dimensions, where each corner lies on its own axis. In the topic-word distribution, each of the "corners" of this triangle represents one topic, and a point within this triangle represents one word. A word near one corner of the triangle is fairly exclusive to that topic, whereas a word closer to the middle of the triangle can be included multiple topics. One hyperparameter (which we will call α), in a sense, "encourages" words to occur either in the corners of the distribution ($\alpha < 1$) or in the center of the distribution ($\alpha > 1$). Likewise, for the document-topic distribution, each corner of the triangle represents one topic, and one point within the triangle represents one document. A document near the center of this triangle is composed fairly equally of all three topics, whereas a document close to one of the corners would fairly exclusively include the topic corresponding to that corner. A second hyperparameter, β, behaves like α and controls a probability distribution which defines the degree to which one document includes a mixture of all the topics. More detailed descriptions of the mathematics and intuition behind LDA can be found in Blei *et al.* [29] and Odden *et al.* [9].

In the context of our research, the corpus of documents is the collection of all student responses for a single question, and a document is a single student response within that corpus. The research we are conducting looks at the extent to which a topic can represent a student idea. In this setting, any given student response would be comprised of a distribution of ideas, with the heterogeneity of the topic-document

distribution controlled by the hyperparameter β. For this preliminary analysis, the values of α and β were chosen automatically by Gensim, which we used to implement the LDA algorithm.

We ran models on the corpus of student responses with the number of topics (k) ranging from three to seven, and computed coherence values (C_v) for each model [30]. The coherence is a measure of "the tendency of the top words in the topic to co-occur"[9], where higher coherence values (closer to 1) describe more distinct topic distributions. In selecting the number of topics to use for our final analysis, were attentive to the topic number which yielded the highest coherence value, and the topic number that appeared to yield the most interesting student ideas. This was done by inspection; two authors examined which topics looked to be the most distinct and/or best matched our instructional experience. In practice, the models with higher coherence also appeared to produce more distinct or instructionally-relevant topics. For this analysis, we chose to use a five-topic model, which yielded the highest coherence value.

Because LDA relies on an initial randomization, it is important to note that between random seeds, there is some run-to-run variation in the different topics the model converges on. The topics presented in this paper are chosen from just one particular run of the model; results may vary if a random seed for the algorithm is not specified.

D. Representative Responses

Because LDA modeling assigns a weight for each topic within the document, we were able to examine the documents with the highest weights for each topic. We considered the documents with the highest weight for a given topic to be the best examples of that topic. Without displaying the student responses, we would need to infer student ideas solely from the words in the topic and their respective weights. Examining representative responses allowed us to ground our interpretation of topics-as-ideas in real student responses that are "closest" to each topic.

We used the three most representative responses for each topic in conjunction with the topic words and their respective weights to characterize students' ideas about the given circuits question. In this secondary, qualitative analysis of the LDA model, we (a) examined the highest-weighted words in a topic, (b) picked out phrases in the representative responses that contained these words, and (c) synthesized those phrases into coherent physics ideas guided by our "professional vision" as physics instructors [31]. We found this final interpretive step crucial for assigning disciplinary meaning to the model's topics.

III. RESULTS

Here we present the topics characterized by the model and representative responses from each topic, and we discuss how we used these results of the modeling process to characterize five distinct student ideas about circuits. The topics produced by the model are shown in Table I. This five-topic model yielded a coherence score of 0.4625. Table II shows some phrases from the most representative student responses for each topic.

TABLE I. Top words in topics with weights

Topic	**Words in Topic** (with weights)
1	water (.090), large (.053), flow (.053), think (.044), energy (.042), pressure(.040), like (.028), great (.026), push (.023), electron (.023)
2	electron (.074), force (.057), great (.044), high (.041), big (.038), mean (.033), faster (.031), potential (.027), push (.023), think (.022)
3	increase (.053), high (.052), charge (.050), electron (.039), great (.033), battery (.031), lead (.029), flow (.027), mean (.027), potential_difference (.026)
4	battery (.081), circuit (.069), power (.069), bulb (.048), increase (.043), brighter (.036), think (.029), push (.028), lead (.024), double (.022)
5	increase (.163), resistance (.083), circuit (.048), constant (.041), equation (.027), mean (.024), bulb (.024), ir (.024), ohm_law (.022), change (.022)

TABLE II. Selected phrases from most representative responses (Topic words in **bold**)

Topic	**Representative Student Responses**
1	"**water flows** through a river **like** current **flows** through circuits" "if there is more potential (or, less accurately, **pressure**)...more **electrons** will **flow**"
2	"the more **potential**...means the more **force** there is **pulling** [the **electron**] to where it wants to be" "more voltage...results in a harder **pull**, and therefore the electrons speed up more...**faster** electrons will result in a **higher** current"
3	"if the voltage of the **battery increases**...**potential difference**...**increases**...the **battery** must push more charge through it to maintain the voltage" "a **greater** magnitude electric field...**increases** the amount of **charge** that passes through a...circuit...**increasing** the current"
4	"second **battery**...**doubles** the voltage...**increasing** the current **increases** the **power**" "a **brighter bulb** represents more **power** and therefore more current"
5	"rearranging the equation $V = \boldsymbol{IR}$...in terms of the **equation**...a change in one value affects the other...if R is **constant**, I must **increase**."

Our qualitative analysis of the five topics given by the LDA model produced the following distinct student ideas:

1. Current flows through circuits like water through a river or pipe.
2. Voltage results in a force which "pulls electrons."
3. Voltage is potential difference, which pushes charges. More V means more charges pushed, so more I.
4. Batteries increase the power in a circuit. More power

means brighter.

5. Voltage is proportional to current. With resistance held constant, increasing the voltage increases the current.

For this particular set of responses to the circuits question, the LDA algorithm produced topics that we were able to interpret as distinct student ideas. Topic 1 was a particularly distinct and stable topic; every run of the algorithm (regardless of number of topics) produced some topic related to the flow of water. Though Topics 2 and 3 were both related to potential and current, we noted that Topic 2 related voltage to a force on electrons causing them to speed up, whereas representative responses for Topic 3 focused more on a higher voltage "pushing more charges." While subtle, this minute linguistic difference was significant enough for LDA to distinguish between the two ideas. Topic 4 refers to the idea that "batteries provide power," including colloquial notions of power (i.e. a "power outage") and more technical definitions (e.g., $P = IV$). Topic 5 refers directly to Ohm's Law; responses representative of this topic use the equation itself to explain why increasing the voltage should increase the current.

IV. CONCLUSION/DISCUSSION

The research questions guiding this investigation were: *To what extent can LDA be used to characterize patterns in student thinking?* and *Can we produce a time-saving method using LDA that yields instructionally-relevant student ideas?* In this preliminary study, LDA was used to characterize five distinct student ideas about a simple circuit, and these five topics were recognizable to us as physics ideas that were relevant for the question at hand. More specifically, the model on its own produced five distinct topics with relevant physics words, and researchers were readily able to interpret the combined output of topics and representative responses as disciplinarily meaningful ideas.

A primary goal of this analysis was to propose a method to streamline studies of common student ideas. Our results suggest that LDA may support researchers in more efficiently characterizing student ideas, but it does not remove the need for researcher interpretation. To inform further development of a semi-automated method for characterizing student ideas about physics topics, we have discussed some ways in which researchers' decisions about the LDA model (e.g., choices about α and β values or the number of topics) can affect the topics it yields. We have also described the methodological steps taken to interpret instructionally-relevant student ideas from algorithm-generated topics. These choices about the model and the interpretive steps that follow suggest a framework for how future work can use LDA to characterize student ideas.

Future work could refine the use of LDA for characterizing student ideas by investigating how the hyperparameters α and β can be chosen to produce the most instructionally-relevant topics. There is no theoretically-based method choosing the "best" values for α and β for the LDA algorithm [32]. Rather, these hyperparameters should be chosen based on the research goals and aims. In our model, we allowed these parameters to be chosen automatically. We are curious whether lower values of α and/or β might produce more distinct, easy-to-interpret topics. Future work could include an exploration of how tuning the various model parameters affects the coherence scores or the perceived instructional relevance of the model output.

Some limitations of this approach for categorizing student ideas include the need for a large sample and from the fact that LDA looks specifically for patterns in text. Typically, LDA is best suited to a corpus containing upwards of $N = 1000$ documents, preferring a large amount of small documents over a small amount of large documents [9]. With $N = 483$ student responses, the run-to-run topic variation was not incredibly high, but some runs of the model were noticeably different. As a test, we ran this model on a data set of $N < 200$ responses and we found that the run-to-run variation in topics was much higher and the topics were less distinct and more difficult to interpret. This suggests to us that in order for LDA to be an effective tool, roughly 500 or more student responses are necessary. For smaller datasets (e.g., $N < 300$ responses) topics are more difficult to interpret from the LDA model and hand-coding is less time consuming, and therefore may be preferable.

Students' written responses to conceptual physics questions typically vary significantly in terms of the visual and mathematical representations they use, their spelling and word choices, and their length. A researcher manually parsing through students' written responses can interpret the meanings of various representations, synonyms, and misspellings. Electronically collected responses do not always allow students to include diagrams or other visual representations of their thinking, and in any case LDA cannot interpret visual representations. LDA does not match physically equivalent equations or expressions, and thus may miss when words or topics are nearly identical from a physics perspective. Additionally, misspelled words lose their meaning in the LDA model. Lastly, student homework responses vary from just a few words in length to several sentences forming a paragraph or two. LDA treats these as having equal footing, though an instructor might expect that longer responses include more distinct ideas. Each of these issues limits the extent to which the LDA algorithm captures the full meaning of a student's response. Future work could include more data cleaning and pre-processing, including spell-checking and splitting responses up into one-sentence chunks which act as documents; this may improve model stability and usefulness for characterizing distinct, common ideas.

The code used for this analysis, and other case studies using questions about heat & temperature and waves, are available on GitHub as Jupyter Notebooks [33].

ACKNOWLEDGMENTS

We are grateful to Lauren Bauman, Amy Robertson, and Paula Heron for their feedback on a draft of this manuscript. This work was supported in part by NSF grants 1914603 and 1914572.

[1] I. Portugal, P. Alencar, and D. Cowan, Expert Systems with Applications **97**, 205 (2018).
[2] J. Ni, Y. Chen, Y. Chen, J. Zhu, D. Ali, and W. Cao, Applied Sciences **10**, 2749 (2020).
[3] E. D. Liddy, Encyclopedia of Library and Information Science, 2nd Ed. (2001).
[4] I. Vayansky and S. A. Kumar, Information Systems **94**, 101582 (2020).
[5] E. Linstead, C. Lopes, and P. Baldi, in *2008 seventh international conference on machine learning and applications* (IEEE, 2008) pp. 813–818.
[6] S. Moro, P. Cortez, and P. Rita, Expert Systems with Applications **42**, 1314 (2015).
[7] J. Munsell, N. S. Rebello, and C. Rebello, in *Physics Education Research Conference 2021*, PER Conference (Virtual Conference, 2021) pp. 295–300.
[8] C. Zabriskie, J. Yang, S. DeVore, and J. Stewart, Physical Review Physics Education Research **15**, 020120 (2019).
[9] T. O. B. Odden, A. Marin, and M. D. Caballero, Phys. Rev. Phys. Educ. Res. **16**, 010142 (2020).
[10] T. O. B. Odden, A. Marin, and J. L. Rudolph, Science Education **105**, 653 (2021).
[11] L. C. McDermott, P. S. Shaffer, and University of Washington Physics Education Group., *Tutorials in Introductory Physics* (Pearson Learning Solutions, 2012).
[12] R. M. Goertzen, E. Brewe, L. H. Kramer, L. Wells, and D. Jones, Physical Review Special Topics-Physics Education Research **7**, 020105 (2011).
[13] L. S. Shulman, Educational researcher **15**, 4 (1986).
[14] M. R. Stetzer, P. van Kampen, P. S. Shaffer, and L. C. McDermott, American Journal of Physics **81**, 134 (2013).
[15] L. C. McDermott, American Journal of Physics **69**, 1127 (2001).
[16] L. C. McDermott and E. F. Redish, American Journal of Physics 10.1119/1.19122 (1999).
[17] A. A. diSessa, Cognition and Instruction **10**, 105 (1993).
[18] D. Hammer, American Journal of Physics **68**, S52 (2000).
[19] D. Hammer, A. Elby, R. E. Scherr, and E. F. Redish, Transfer of learning from a modern multidisciplinary perspective **89** (2005).
[20] L. M. Goodhew, A. D. Robertson, P. R. L. Heron, and R. E. Scherr, Phys. Rev. Phys. Educ. Res. **15**, 020127 (2019).
[21] L. C. Bauman, J. Corcoran, L. M. Goodhew, and A. D. Robertson, in *2020 Physics Education Research Conference Proceedings* (2020) pp. 57–62.
[22] P. V. Engelhardt and R. J. Beichner, American Journal of Physics **72**, 98 (2004).
[23] L. M. Goodhew, A. D. Robertson, P. R. Heron, and R. E. Scherr, Physical Review Physics Education Research **17**, 10137 (2021).
[24] R. Řehůřek, P. Sojka, *et al.*, Retrieved from genism. org (2011).
[25] W. McKinney *et al.*, Python for high performance and scientific computing **14**, 1 (2011).
[26] G. Van Rossum, *The Python Library Reference, release 3.8.2* (Python Software Foundation, 2020).
[27] Y. Zhang, R. Jin, and Z.-H. Zhou, International Journal of Machine Learning and Cybernetics **1**, 43 (2010).
[28] J. Lin, Department of Mathematics and Statistics, Queens University (2016).
[29] D. M. Blei, A. Y. Ng, and M. I. Jordan, Journal of machine Learning research **3**, 993 (2003).
[30] M. Röder, A. Both, and A. Hinneburg, in *Proceedings of the eighth ACM international conference on Web search and data mining* (2015) pp. 399–408.
[31] C. Goodwin, American Anthropologist **96**, 606 (1994).
[32] C. P. George, H. Doss, *et al.*, J. Mach. Learn. Res. **18**, 5937 (2017).
[33] https://github.com/jonmgeiger/honors-project.