Journal of Science Education and Technology
https://doi.org/10.1007/s10956-020-09856-2

How Does Augmented Observation Facilitate Multimodal Representational Thinking? Applying Deep Learning to Decode Complex Student Construct

Shannon H. Sung[1] · Chenglu Li[2] · Guanhua Chen[3] · Xudong Huang[3] · Charles Xie[1] · Joyce Massicotte[3] · Ji Shen[4]

© Springer Nature B.V. 2020

Abstract
In this paper, we demonstrate how machine learning could be used to quickly assess a student's multimodal representational thinking. Multimodal representational thinking is the complex construct that encodes how students form conceptual, perceptual, graphical, or mathematical symbols in their mind. The augmented reality (AR) technology is adopted to diversify student's representations. The AR technology utilized a low-cost, high-resolution thermal camera attached to a smartphone which allows students to explore the unseen world of thermodynamics. Ninth-grade students ($N = 314$) engaged in a prediction–observation–explanation (POE) inquiry cycle scaffolded to leverage the augmented observation provided by the aforementioned device. The objective is to investigate how machine learning could expedite the automated assessment of multimodal representational thinking of heat energy. Two automated text classification methods were adopted to decode different mental representations students used to explain their haptic perception, thermal imaging, and graph data collected in the lab. Since current automated assessment in science education rarely considers multilabel classification, we resorted to the help of the state-of-the-art deep learning technique—bidirectional encoder representations from transformers (BERT). The BERT model classified open-ended responses into appropriate categories with higher precision than the traditional machine learning method. The satisfactory accuracy of deep learning in assigning multiple labels is revolutionary in processing qualitative data. The complex student construct, such as multimodal representational thinking, is rarely mutually exclusive. The study avails a convenient technique to analyze qualitative data that does not satisfy the mutual-exclusiveness assumption. Implications and future studies are discussed.

Keywords Heat transfer · Representational thinking · Augmented observation · Bidirectional encoder representations from transformers (BERT) · Transfer learning · Automated text classification

Introduction

External representations display abstract concepts with concrete symbols or analogy (Shen and Confrey 2007; Xie 2011; Namdar and Shen 2015). Representational thinking, which could be operationally defined as the construct encoding how one mentally forms conceptual, perceptual, graphical, or mathematical symbols, is an essential precursor to foster proficient math and science practices (Namdar and Shen 2015; Samuelsson et al. 2019; Jitendra et al. 2016; National Research Council 2008; Next Generation Science Standards Lead States 2013). In the past decade, the advancement of AR technologies has transformed how external representations could foster the representational thinking in the science laboratory (Garzón and Acevedo 2019). Nevertheless, researchers almost always determine the effectiveness of AR by student's improved conceptual understanding or attitude (Chang et al. 2018; Dunleavy et al. 2009; Garzón and Acevedo 2019; Wu et al. 2013; Wu and Puntambekar 2012). Very few science educators studied representational thinking in an AR-assisted laboratory, probably because students often adopt multiple representations at once, and the assessment could be very

✉ Shannon H. Sung
shannon@intofuture.org

[1] Institute for Future Intelligence, 26 Rockland St., Natick, MA 01760, USA

[2] School of Teaching and Learning, University of Florida, 1221 SW 5th Ave, Gainesville, FL 32601, USA

[3] Concord Consortium, 25 Love Lane, Concord, MA 01742, USA

[4] Department of Teaching and Learning, University of Miami, 5202 University Drive, Coral Gables, FL 33124, USA

Published online: 16 September 2020

arduous and time-consuming for human coders (Sung et al. 2020). We resorted to the machine as an apprentice of humans to relieve the constraint of the limited human-labor capacity. The previous label classification in natural language processing (NLP) all assume one code, and only one, to each analysis unit against all available codes (e.g., Ha et al. 2011; Nehm et al. 2012; Liu et al. 2016a). The existing machine learning method would be unsuitable to perform desired automated assessment, because most students innately manipulate multiple modes of representational thinking at the same time (Sung et al. 2020). It is imperative to search for an alternative method that could be customized to meet our research goal in automatizing the coding process. The study is, therefore, guided by the research question:

> *What is the affordance of machine learning in promptly assessing a student's multimodal representational thinking in an AR-assisted lab?*

Representational Thinking Functioning in an AR-Assisted Thermodynamics Lab

Students constantly retrieve information from their cognitive blackbox to explain abstract concepts like thermodynamics (Clark and Jorde 2004; Pathare and Pradhan 2010; Reiner et al. 2000; Wiser and Kipman 1988). In contrast to the thermal vision possessed by some animals, humans need visual aids to see the invisible heat transfer process (Xie 2011). This type of super-sense might facilitate learning by expanding the limited perceptual or visual capacity of students (Moreno 2006). According to the cognitive theory of multimedia learning, both external (i.e., verbal and visual) and internal representations (i.e., tactile perception, mental image) could also be used as one of the multimodal representations to diversify the information input and, thus, increase the cognitive capacity (Mayer and Moreno 2003; Moreno 2006). Therefore, the representational thinking could be understood by how students mentally process the sensory information input, which could be in the form of physical touch, numerical symbols from a temperature probe, verbal signals, etc. (Sung et al. 2020; National Research Council 2008). It is believed that students' adoption of multimodal representational thinking is an essential predecessor for conceptual change and development of scientific modeling (National Research Council 2008). In this article, the multimodal representational thinking could be conceptualized as how one actively processes the multiple external representations of thermal conductivity afforded by the innovative technology (Sung et al. 2020). This assumption is grounded in the cognitive theory of multimedia learning (Mayer 2014), where Mayer suggested that people take in sensory information and actively filter, select, organize, and integrate the sensory input into mental constructs.

Researchers have been dedicated to developing multimodal computerized representations as learning aids to promote normative scientific understanding of heat (Clark and Jorde 2004; Clough and Driver 1985; Donnelly et al. 2015; Lewis and Linn 1994; Pathare and Pradhan 2010; Wiser and Kipman 1988). A smartphone app that incorporates infrared radiation (IR) thermal imaging technology (i.e., SmartIR) could create innovative channels to augment observation, and the external representations could be internalized to help the users to understand heat conduction concepts (Samuelsson et al. 2019). We operationally defined multimodal representational thinking of heat as one's ability to mentally portray the multiple external representations (Wu and Puntambekar 2012), such as symbols, graphs, visual aids, and even thermal sensation input as meaningful information for future retrieval. More specifically, we hypothesized that the internal mental representations were externalized in the form of text-based explanation when multimodal representational thinking is exercised during science inquiry. This method deviates from studies that adopt representational competency (e.g., Magana and Balachandran 2017) or representational fluency (e.g., Moore et al. 2020) frameworks that were applied to directly encapsulate and assess the representational quality of students' artifacts. We treated the multimodal representational thinking more in line with how students embodied multiple representations (Walsh et al. 2020) and whether they apply and transform the cognitive construct into textual representation in explaining their thermal sensation and symbolic artifacts (Sung et al. 2020).

Researchers in cognitive sciences, mathematics, and science disciplines are dedicated to studying how external representations are embodied and internalized into the form of representational thoughts (Ainsworth 2008; Jitendra et al. 2016; Walsh et al. 2020; Mayer 2014; Moore et al. 2020; Walsh et al. 2020; Wu and Puntambekar 2012); however, to what extent does students incorporate their mental representations in explaining perception and interpreting data collected during the laboratory activity remains scarce (Sung et al. 2020), not to mention automated coding.

Applying Machine Learning in Automated Assessment

Human coders rely heavily on a manual process accompanied by constant comparison and contrast of their codes with others to reconcile disagreement and to cross-validate their codes until reaching satisfactory interrater reliability (Bailyn 1977; Carbó et al. 2016; Este et al. 1998; Grolemund and Wickham 2014; Patton 2005; Tanana et al. 2016). The systematic training of human coders is labor-intensive and costly (Beggrow et al. 2014; Blank 2004; Liu et al. 2016a; Nehm et al. 2012; Tanana et al. 2016). Researchers resorted to machine scoring tools, such as Computer-Assisted Scoring (CAS) (Beggrow

et al. 2014), c-rater-ML (Liu et al. 2016a), SPSS Text Analytics (SPSSTA) (Haudek et al. 2012), Summarization Integrated Development Environment (SIDE) (Ha et al. 2011; Nehm et al. 2012), and so forth, to automate the scoring process. Among many machine learning techniques, natural language processing (NLP) is commonly applied in assessing open-ended responses (e.g., Luo and Litman 2016), interview transcripts (e.g., Beggrow et al. 2014; Tanana et al. 2016), and other text artifacts (e.g., Krithika and Narayanan 2015). Some NLP approach was completed by combining the baseline feature (Luo and Litman 2016) and by applying clustering analysis on scoring short text responses (Beggrow et al. 2014; Zehner et al. 2016) and essays (Haudek et al. 2012; Mehmood et al. 2017), while other NLP algorithms were adopted to extract linguistic features from students' text responses (Allen et al. 2016; Goodfellow et al. 2016; Zehner et al. 2016). Since NLP is a mature method of automated assessment, it has been widely applied in grading scientific explanations (e.g., Ha et al. 2011; Liu et al. 2016a; Nehm et al. 2012). The machine learning–based NLP did open up a promising avenue to relieve the laborious task of coding student artifacts, because it could generate reliable scores that are comparable with human coders (Ha et al. 2011; Liu et al. 2016a; Mitchell 1997; Tanana et al. 2016; Zehner et al. 2016). The drawback of the aforementioned tools is that they function under the assumption that the codes are mutually exclusive to each other. In our previous study, we found that participants' language-based representations gave us some hints that the participants have applied multimodal representational thinking in explaining their tactile perception, pictorial thermal imaging data, and graphical artifacts (Sung et al. 2020). However, the process of applying multiple codes in accordance with the coding scheme was time-consuming and labor-intensive. An automated coding algorithm could be the rescue.

Identifying Multimodal Representational Thinking via Multilabel Text Classification

Previous educational research on both automated coding and scoring primarily focused on the assignment of one code to each artifact and assumed that the classification of codes is mutually exclusive (e.g., Beggrow et al. 2014; Ha et al. 2011; Liu et al. 2016a). Despite reported success in applying NLP in automated coding, the limitation of the mutually exclusive coding assumption does not entirely satisfy the requirement in capturing multimodal representational thinking. Alternative machine learning algorithms that could adequately classify the complex cognitive processes with multiple possible codes are called for. Luckily, a machine learning method—multilabel text classification—could serve the purpose (Tsoumakas and Katakis 2007).

In multilabel classification, there are more than two labels for each selection, and there can be more than two labels predicted by the classifier (Tsoumakas and Katakis 2007). The multilabel classification techniques are (1) problem transformation methods and (2) algorithm adaptation methods (Tsoumakas and Katakis 2007). With problem transformation methods, researchers frame multilabel classification problems as one or more single-label classifications or regression problems. Common techniques of problem transformation methods include (a) binary relevance, which builds K binary classifiers for K labels, where K is the total unique count of labels; (b) classifier chains that are built upon binary relevance but also consider the correlation among labels; (c) label powerset, which builds a multilabel classifier that treats every combination of labels as new classes and then solves the problem using multilabel classification approaches; (d) pruned set, which is similar with label powerset but removes labels whose counts are lower than a user-defined threshold; and (e) ensemble methods that build a set of multilabel classifiers with the above problem transformation methods while making the decisions on label predictions with a weighted vote by the multiclass classifiers (Nareshpalsingh and Modi 2017; Zhang and Zhou 2013). Due to its relative simplicity and generalizability, binary relevance is widely adopted for multilabel classification among the problem transformation methods and showed desired results (SpolaôR et al. 2013; Yeh et al. 2017; Zhang et al. 2018a). Feng et al. (2018) used the binary relevance and label powerset methods to build multilabel classifiers and compared their results with multiclass classifiers. The results revealed that multilabel classification with problem transformation methods can yield more accurate results than the alternative classifiers.

For algorithm adaptation methods, researchers modify existing algorithms to handle multilabel classification directly. While there are several methods for algorithm adaptation, most of them are specific for traditional machine learning algorithms such as decision trees and support vector machine (SVM). The traditional machine learning methods are not applicable to process up-to-date algorithms, such as deep neural networks (Nareshpalsingh and Modi 2017). The neural network (NN)-based method is usually adopted to modify the loss function. The NN-based models allow independent probabilities for each label (Nareshpalsingh and Modi 2017), which is by nature similar with binary relevance and classifier chains.

Deep Learning Method and the Mighty Transformer

In recent years, deep learning or deep neural networks have received heightened attention in the field of machine learning. Unlike traditional machine learning approaches, such as SVM, deep learning empowers researchers with more accurate and desired results while requiring less effort on manually

extracting features to optimize model performance (Goodfellow et al. 2016). The advancement of deep learning has drastically improved the fields of speech recognition, visual recognition, object detection, and many other domains (LeCun et al. 2015). The existing educational studies suggest that recurrent neural network (RNN), convolutional neural network (CNN), and more standard machine learning techniques, such as decision trees, are primary methods for automated text classification (Kastrati et al. 2019; Nanaware 2018; Xing and Gao 2018). Researchers started to adopt the transformer mechanism in completing text classification tasks (Vaswani et al. 2017). The transformer is a deep learning architecture for NLP that is substantially optimized in its computation capability compared to other deep learning models, such as RNN and CNN (Vaswani et al. 2017; Zhang et al. 2018b). In addition to computational efficiency, models derived from transformer architecture have performed outstandingly in various NLP tasks such as text classification, reading comprehension, and question-answering thanks to its attention-based mechanism that is capable of capturing nuances in linguistic contexts (Devlin et al. 2018; Li et al. 2018). The transformer-based models that have outstanding performance could also be properly applied to transfer learning, which was not successful for deep learning models such as RNN and CNN in NLP (Mou et al. 2016). Researchers can leverage this layer of flexibility offered by transfer learning to build new models based on existing parameters trained by huge datasets (Devlin et al. 2018).

Models using transfer learning and transformer techniques perform well on natural language inference and paraphrasing (Fedus et al. 2018; Ruder et al. 2019), which are valuable in increasing the accuracy of classification (Liu et al. 2016b). In this study, we proposed the use of the deep learning model—bidirectional encoder representations from transformers (BERT) (Devlin et al. 2018)—to automate the analysis of students' textual responses. First and foremost, BERT performance was outstanding on processing NLP tasks on reputable open-source datasets, including text classification tasks (Devlin et al. 2018). Furthermore, BERT was revolutionary for NLP tasks since it greatly improved on sentence-level learning compared with previous deep learning models (Devlin et al. 2018). Before BERT was created, programmers had been exerting their primary effort in improving the performance for word embedding (Howard and Ruder 2018; Pennington et al. 2014; Peters et al. 2018; Rong 2014). Word embedding is a machine learning technique that assigns words with numbers so that words used in similar contexts denote similar numeric representation. It is advantageous over the traditional bag of word (BOW) method, which simply represents words with their frequency of occurrence in a sentence (Lilleberg et al. 2015). It usually outperforms BOW because word embedding is capable of capturing the *semantic* meanings of words (Lilleberg et al. 2015). However, most of the word embeddings research neglect the *contextual* meanings of words. For example, previous word embeddings could not differentiate the term "pen" which is used for writing, versus a pig "pen" on the farm. Although recent works such as Embeddings from Language Models (ELMo) were able to learn language contexts (e.g., Peters et al. 2018), the nature of learning in such word embeddings was unidirectional (left-to-right or right-to-left), which was suboptimal for performance and might be problematic to adopt in further transfer learning techniques (Devlin et al. 2018). BERT, on the other hand, was able to perform NLP tasks based on both left and right contexts due to its original language representation learning that simultaneously considers bidirectional contexts (Devlin et al. 2018). Thanks to the bidirectional nature of deep learning, BERT helped researchers to analyze text data with higher precision. We take advantage of the affordance of text classification in assessing student's multimodal representational thinking in the AR-assisted lab.

To the best of our knowledge, this preliminary study would present the first research attempt to apply multiple-labeling technique in coding students' multimodal representational thinking that is externalized in the form of written representations. The complex construct is believed to be facilitated by the mobile AR technology (Sung et al. 2020). We compared two machine learning techniques: one is a commonly used machine learning method (i.e., SVM) and the other is the most recent deep learning technique (i.e., BERT). Specifically, we applied these two machine learning techniques to extract the cognitive processes associated with the multimodal representational thinking found in student lab reports.

Methods

Our team developed a smartphone IR application (i.e., SmartIR) that provides (1) color heat map, (2) virtual thermometer, and (3) temperature–time graph features as three modes of external representations to understand heat transfer. These augmented observation features were designed to support representational thinking of users (see Fig. 1) (Sung et al. 2020). Specifically, the vision is augmented to see heat transfer, and the tactile thermal perception is augmented with a digital thermometer to transduce sensory data to an exact temperature reading. The temperature–time graph shows that heat transfer is a "process" and the temperature of a particular spot receiving thermal energy would change accordingly with time. In summary, the smartphone application (i.e., SmartIR) could equip learners with spatial, transductive, and temporal augmented thermal perception (see Fig. 2), to "see" heat transfer. The cognitive functioning of augmented observation features on SmartIR is introduced in the captions of Fig. 2 (see Sung et al. 2020 for more details about the development of SmartIR).

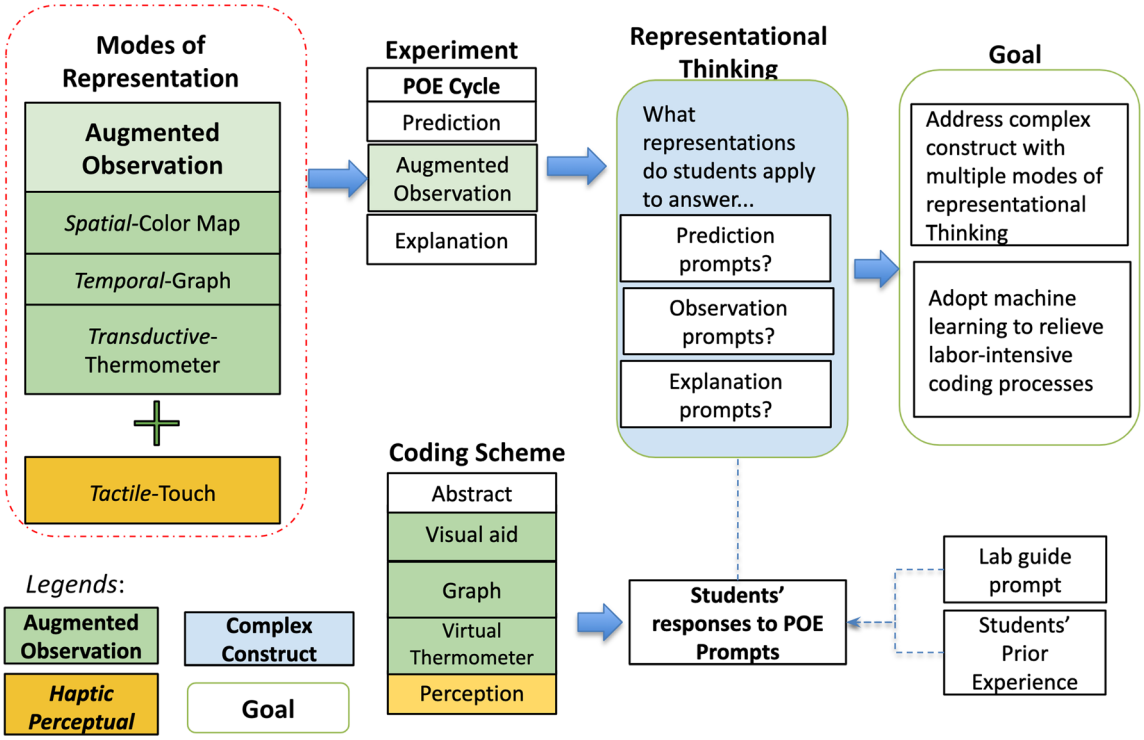


Fig. 1 The figure depicts the research design and theoretical framework of the source of representational thinking. The components in red-dashed box indicate different modes of representations during the AR-assisted heat conduction lab. The four modes of representations (three external representations (green box) and one perceptual representation (orange box)) adopted during the augmented observation phase of the prediction–observation–explanation (POE) cycle could be used to activate fluid representational thinking. The coding scheme, the sources for responses, and the goal of this study were also included in the research outline. Machine learning was adopted to mitigate the researchers' effort in determining multimodal representational thinking

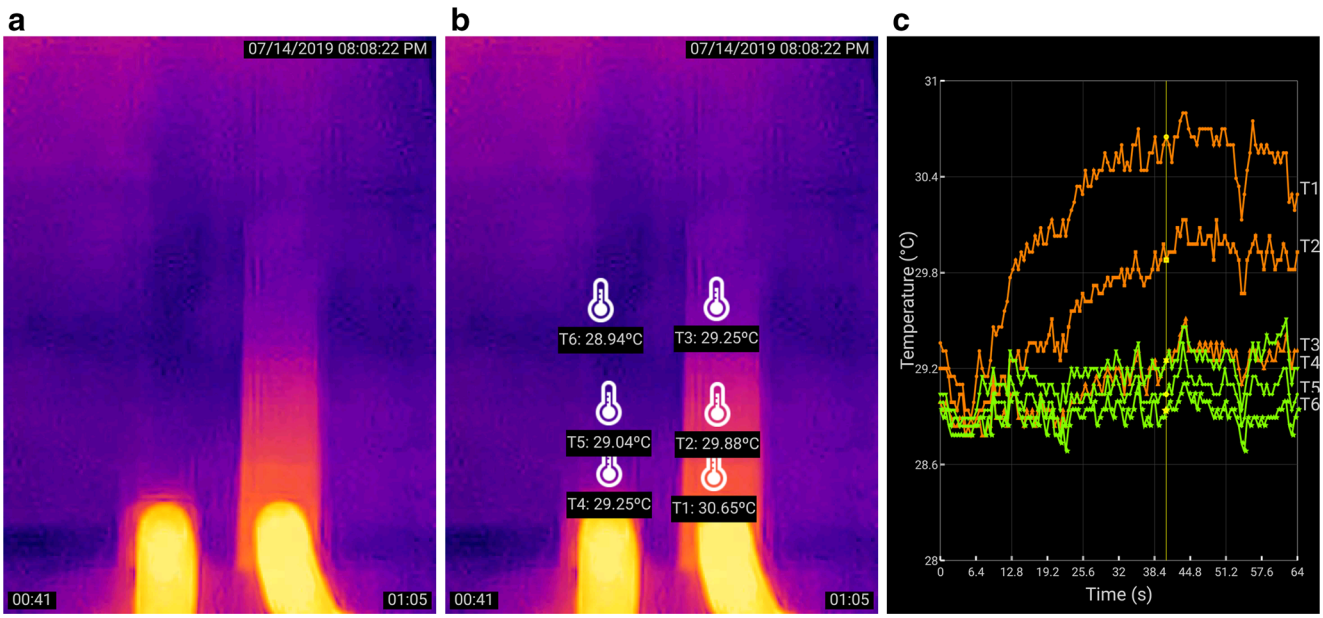


Fig. 2 **a** Spatial augmentation: The *color heat map* augments vision to leverage understanding of energy transfer. **b** Transductive augmentation: The *virtual thermometer* augments observation via transducing the differential pixels on the heat map into respective temperature readings. **c** Temporal augmentation: The temperature–time graph augments observation by highlighting the change of temperature over time

Module Design

We adopted the prediction–observation–explanation (POE) design pattern (e.g., Ebenezer et al. 2010). The "Two-Thumbs Up" module used in this study focuses on heat conduction and thermal perception of heat transfer (see Appendix Table 3 for activities and the POE question prompts embedded in the module). The activity scaffolded students' representational thinking by means of incorporating their thermal sensation as well as the augmented observation features. Student's multimodal representation was elicited when responding to the questions included in the module.

Participants and Context

Ninth-grade students from three public schools in the northeastern USA ($N = 314$) engaged in a heat conduction experiment using the augmented observation tools developed for SmartIR application (Fig. 2a–c; Sung et al. 2020). SmartIR could be used to address science standards related to thermodynamics to learn "*the **temperature** of a substance changes when thermal energy is transferred **from** or **to** a sample*" (LP E04) (Next Generation Science Assessment Task Collections 2019). The IR thermal imaging technology enables students to experience thermal vision to actively observe the heat transfer process, which facilitated augmented observation.

Data Collection and Analysis

Questions in the "Two-Thumbs Up" module were designed to elicit students' representational thinking to explain (1) the thermal perceptions of both wooden and metal rulers before and after they touch them and (2) their thermal sensation of their thumbs (see Appendix Table 3). The questions prompted students to reflect on their augmented observations enabled by the SmartIR. Student's representational thinking in the AR-assisted lab was inferred by the answers they wrote in the lab report in an electronic format when students responded to POE prompts. Each student's text response to every question was considered as one coding unit. Four researchers discussed the responses of one class in search of the evidence supporting the themes and reached interrater reliability of 0.82 after coding 43.37% of a total of 572 responses (see Sung et al. 2020 for more details). Five main categories were summarized as a result of the iterative human coding processes: (1) ideas focusing primarily on haptic perception (P), (2) abstract ideas (A), (3) thermometer reading (R), (4) color heat map visualization (V), and (5) temperature–time graph (i.e., temperature against time graph, $T(t)$) (G) (see Appendix Table 4 in the supplement for the detailed coding scheme for human raters). Multiple codes might be assigned to the same coding unit, except for those responses with off-task or abstract ideas (see Appendix Table 4 for the coding scheme to train NLP).

After coding, 416 responses were single labeled (73%), 111 responses were multilabeled (19%), and 45 responses (8%) were coded as *no response* (NR) since their content was empty or meaningless. Table 1 shows the frequency and examples of each coding scheme.

Text Classification Pipeline

Figure 3 shows the steps we took to build and evaluate the multilabel text classifiers with Python packages *SpaCy*, *Scikit-learn*, and *PyTorch*: (1) We conducted data preprocessing to normalize texts, which includes white space stripping on the start and end of each response and transforming responses into lower-case forms (Pranckevičius and Marcinkevičius 2017; Nayak and Natarajan 2016); (2) we extracted linguistic features from responses which allowed us to potentially enhance the training of models. Although the extraction of such features was controversial (Schenk et al. 2016), significant improvement on performance was observed (Pennacchiotti and Popescu 2011; Pla and Hurtado 2014); (3) we trained multilabel classifiers of SVM and BERT with plain texts and texts with linguistic features, respectively; (4) we used common metrics to evaluate the performance of multilabel classifiers; and (5) we chose the best-performing model and used it to predict the rest of the unlabeled data ($n\_unlabeled = 648$). Details of each step are given in the following sections.

Feature Extraction

We used Python SpaCy to build two linguistic features: part-of-speech (POS) and named entity recognition (NER). POS tags describe the grammatical functionalities of words such as verbs, adjectives, adverbs, and punctuations. NER tags categorize information extracted from words such as names of people, organizations, and locations. Compared with POS, NER tags are more informative, yet less common. Not every response would contain information categorized by NER. Three new datasets were created where preprocessed plain texts were linked with POS, NER, and POS + NER features, respectively. For models of BERT, we modified their embedding layers to treat POS and NER tags as special tokens because BERT's tokenizers would otherwise treat those linguistic features as unknown words and thus affect performance.

Model Training and Performance Measures

In order to evaluate the performance of BERT thoroughly, we also used the traditional machine learning algorithm SVM to provide a base performance. We chose SVM because of its reported desirable performance on multilabel classification with binary relevance

Table 1 Frequency and example of each coding scheme

Code	Examples	Count
P	After we touch them for a minute, the metal will heat up more because it will absorb the heat from our hand faster. (A11^, P2*) I think that the metal ruler will feel cooler because it releases heat better than the wooden ruler. (C9, P1) The metal will feel colder because metal retains far less heat than wood does. (G12, P1)	148
V	The place where the thumb had touched the metal ruler was much higher than the place the thumb had touched the wooden ruler because the metal ruler retained the thermal energy much better as it is a better conductor. (A10, E3) The metal conducts heat better than the wood, which is why the heat was able to travel farther. (G14)	111
A	The will return to the temperature they were before (A11, P3) The metal one would absorb heat more rapidly than the wood one, since metal is a conductor while wood is an insulator. (G7, P2)	59
G	In the graph it shows how the longer we had the thumbs on them the hotter they got. The spike in the middle is the temperature of the thumbs. (A12, Og2) It's demonstrating a spike in thermal energy after a certain period of time that goes down after another period of time. (G8, Ob)	54
R	The pattern shown on the image is two rulers, one metal, one wood, that are both at room temperature. (A7, Oa)	44
P, V	The fingers have lost heat because the heat has flowed into the two rulers. (A4, Od) The heat from my thumbs traveled up the ruler. The wooden ruler barely got warmer except for right under where my thumbs were. (A6, Ob)	54
R, V	The pattern shown in this image is that the rulers are the same color as the foam board behind them. This shows that the rulers are at the same temperature as the board, which is room temperature. This can also be shown by the little thermometers on the rulers because they both show relatively the same temperatures. (A3, Oa) The rulers are at the same temperature as in (b) because the rulers are still heated, the heat source is just removed so the temperature of both rulers will not increase. (G13, Oc)	36
R, P	The thermometer read warmer as they got closer to the thumbs because the thumbs were the heat source and the heat energy was traveling down the ruler through conduction. (G13, E2)	7
P, V, R	One thumb was cooler than the other after touching two rulers, because it gave away more heat to the metal ruler. The metal ruler started out with a cooler temperature but is a conductor, which is why it took more heat from the fingers. While the wooden ruler was the opposite which is why the temperature of the finger that touched the wood was warmer, and the finger that touched the metal was cooler. (G6, E4)	4
G, P	The temperature of the thermometers are heating up with the warmth of my fingers, so the bottom ones are the thermometers on the metal ruler. (A6, Og1)	4
G, R	Both rulers decrease in temperature very quickly. If you look at the temperature, the thermometer is 29.93 °C, while the wooden ruler is 30.07 °C. The metal ruler lost heat faster than the wooden ruler did. (E4, Oc)	2
R, P, V	The metal ruler is increasing temperature with energy flowing from my fingers to the ruler. The wood ruler is darker and has a lower temperature than the metal ruler does. (A7, Ob)	2
P, A	The place where our fingers touch the rulers for longest will be the hottest and the heat will spread from there. The metal ruler will have been heated more thoroughly because it is a better conductor of heat. (A9, P2)	1
V, G	When we put our thumbs on it, the temperature of each ruler increased. The metal increased temperature faster and spread to more of the ruler than the wood did. (A11, G2)	1
NR	No response, blank	45

^Student ID

*POE prompt number

method (Sun et al. 2017; Xu 2011). Python packages PyTorch and Scikit-learn were used to test two types of supervised machine learning algorithms: transfer learning with BERT and SVM. In order to build multilabel classifiers, we adopted binary cross-entropy with logits in BERT's loss function, which allowed us

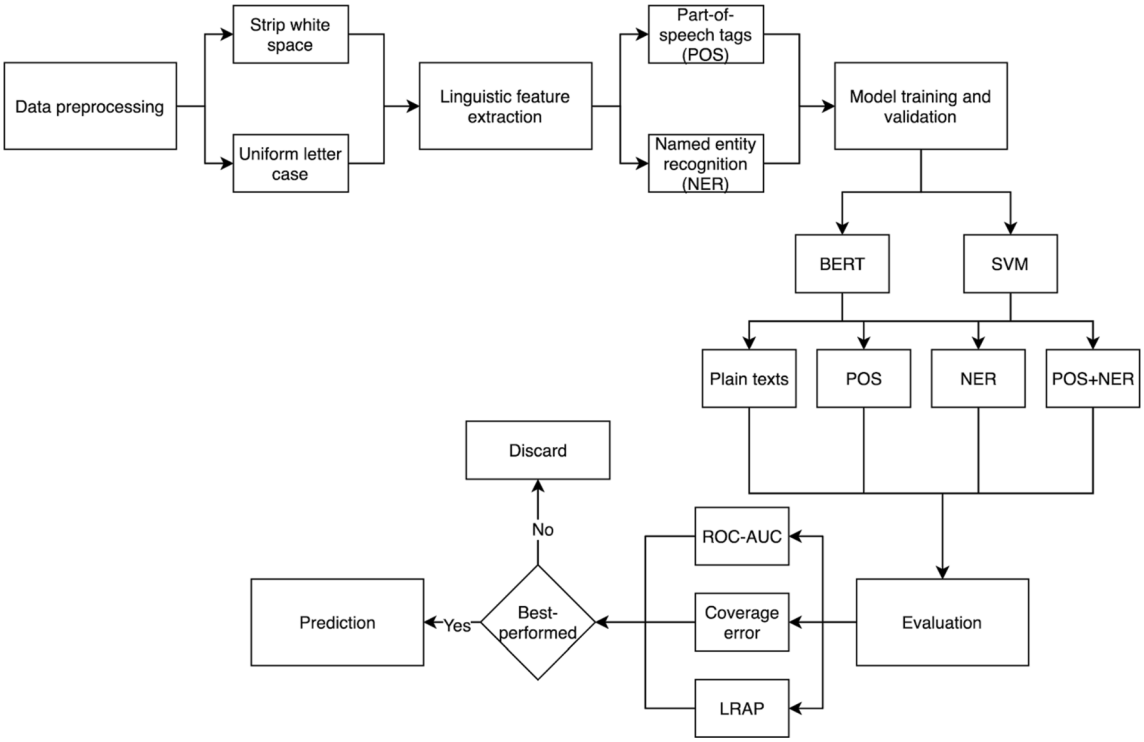


Fig. 3 Block diagram demonstrating data processing flow of BERT and SVM. The figure depicts the index adopted to evaluate prediction accuracy

to assign independent probability ranging from 0 to 1 for each representational thinking label. In terms of SVM, we used the binary relevance method to train six classifiers (P, A, R, V, G, and NR) independently to achieve the goal of multilabel classification.

Depending on which algorithm it adopts, the labeled dataset was split differently in the training process. For SVM, 80% of the dataset was sampled randomly for 5-fold cross-validation and hyperparameter tuning, while the rest was used for testing. For BERT, a majority of the dataset (60%) was used for training, while the rest 40% was evenly divided for validation and testing. Specifically, the labeled data used to train the BERT model was to detect features and classify lab report responses. In the training of BERT, cross-validation with human codes was not performed because the BERT model has 12 layers with 110 million parameters in the training process, and such deep learning models are usually much more expensive to train than traditional machine learning models. Thus, it would be impractical to iteratively cross-validate the training result with human coders. Instead, we used 20% of the labeled data that were not included in the training set to validate the BERT models. The rest of the 20% served as testing data to compare the labels generated by the BERT model and the labels given by human coders, which were the same used for testing in SVM. Metrics received from the testing dataset can serve as an indicator on the model's reliability as well as validity. The testing dataset was labeled by the researchers, thus serving as the golden truths, and the model does not see the content in the testing dataset during training. Therefore, achieving a good performance on the testing dataset indicates a good reliability and validity.

Model Performance Evaluation and Prediction

After model training, we used four metrics to evaluate models that are commonly used for multilabel classification: (1) receiver operating characteristic curve–area under the curve (ROC–AUC), (2) one-error, (3) coverage error, and (4) ranking loss (Alalga et al. 2016; Wu and Zhou 2017; Sarker et al. 2013) (see Fig. 4). The *ROC* curve is a probability curve plotted with false-positive rates in the *x*-axis and true-positive rates in the *y*-axis, and *AUC* is the area under the ROC curve. The *ROC–AUC* score refers to AUC and measures how well a model can discriminate between classes, and the higher the AUC, the better a model predicts true negatives as negative and true positives as positive (Fan et al. 2006). In our case, a high ROC–AUC score means that, for each mode of representational thinking, a model can accurately decide if a response externally represents a certain mode of thinking (e.g., haptic perception (P)) or not.

One-error measures how many times the label with the highest prediction probability does not correspond to that assigned by the human (Tsoumakas et al. 2009). The lower the one-error score, the better a model's predictions are aligned with human-coded responses in terms of predictions with the highest probabilities. For example, a one-error score of 0.1 means, on average, 90% of the time (i.e., 1–0.1 = 0.9) the predicted label with the highest probability is indeed one of the label(s) humans have assigned to student responses. *Coverage error* indicates the average discrepancy between the prediction-probability label ranking versus the human-assigned labels of a response (Tsoumakas et al. 2009). The best value of coverage error is the average number of human-assigned labels in all responses, and we want the model to produce a number close to that average number. *Ranking loss* measures the average fraction of incorrectly ordered label pairs of responses, and the perfect score is 0 (Tsoumakas et al. 2009). Conceptually, repeating in each response, we treat human-assigned labels with values of 1 and otherwise 0 and then sort these labels in a descending order. Then we iteratively sorted labels in pairs, say $label_p$ and $label_v$, meaning the response is labeled as P or V. For each iteration, these two labels are ranked based on prediction probability and their rankings. If the rank of $label_p$ is smaller or equal to that of $label_v$, then we accumulate the loss by 1 because this scenario is deemed as reversely ordered. The accumulated loss will then be divided by the product of the number of 1s and 0s, namely the product of cardinalities between positive and negative cases. Iterating through every response, we will add up the loss and eventually divide the summed loss by the number of responses. That being said, a label ranking of 0.1 means on average 10% label pairs are reversely or incorrectly ordered.

Upon finishing evaluation, we selected the best-performed model that achieved outstanding performance on these four metrics to predict the rest of the unlabeled responses ($n\_unlabeled = 648$). In order to predict the multimodal representational thinking in the multilabel manner, we needed to select a probability threshold to determine above which value the applicable label(s) would be assigned to a response. This step closely resembles the decision-making processes behind human coders. To achieve the purpose of multilabel, we used the SCut method to find an optimized threshold (Al-Otaibi et al. 2014). To be specific, we first generated prediction probabilities of each label in the testing dataset, then we searched in the threshold space within 0 to 1 to find a value that maximizes the micro F-1 score. The micro F-1 score is a value describing sensitivity and specificity weighted by label sample sizes.

Results

Multilabel Classification

Table 2 shows the results of models using BERT and SVM. The BERT model trained with plain texts achieved the best performance on all four metrics, and its performance is significantly better than that of SVM. The high ROC–AUC score of 0.943 from the BERT model suggests its outstanding capability on differentiating between classes. Figure 4 shows the best-performed model of BERT outperforms the best-performed SVM on all labels' AUC. The low one-error of 0.225 shows on average 77.5% (1–0.225) of the top-ranked label by the BERT model is one of the ground true labels. The coverage error of 1.667 indicates the average discrepancy between the label ranking in order to capture all true labels is 1.667, which is closer to the average number of true labels from responses (i.e., 1.234), than SVM. The ranking loss of 0.071 suggests the average percentage of incorrectly ordered label pairs of responses is 7.1%, which means that most of the label pairs are correctly ordered (92.9%). It is interesting to see that adding linguistic features to BERT did not help with its performance and even backfires, which aligns with the findings of Schenk et al. (2016) that deep learning might not benefit from manual feature engineering. On the other hand, SVM achieved the best result when data were augmented with POS features, which resonates with the finding from Kovanović et al. (2014). Since BERT with raw texts was best-performed among all the models, we applied the SCut method on it in search for the probability threshold to accept labels, and the result suggested that a value of 0.272 achieved the best result on the testing dataset (see Fig. 5).

Despite the complex nature of deep learning, we conducted Local Interpretable Model-Agnostic Explanations (LIME) (Ribeiro et al. 2016) to visualize random responses for the observation question to showcase how BERT extracts important features and predicts the representational thinking labels. LIME offers local explanations on models which is only applicable to specific responses. While understanding what a deep learning model has learnt over all responses is possible through methods such as Shapley Additive Explanations (SHAP) (Lundberg and Lee 2017), such methods usually have limited support on deep learning frameworks and thus are challenging to apply. Conceptually, LIME explains how a model weighs words of a specific response by changing texts in the response (e.g., words removal or duplication) and building an explainable linear model with the modified responses. The predictions on modified responses of a blackbox model will serve as the ground truth for LIME's linear model. The explanations from the linear model were reported to be trustworthy explanations from the blackbox model (Ribeiro et al. 2016).

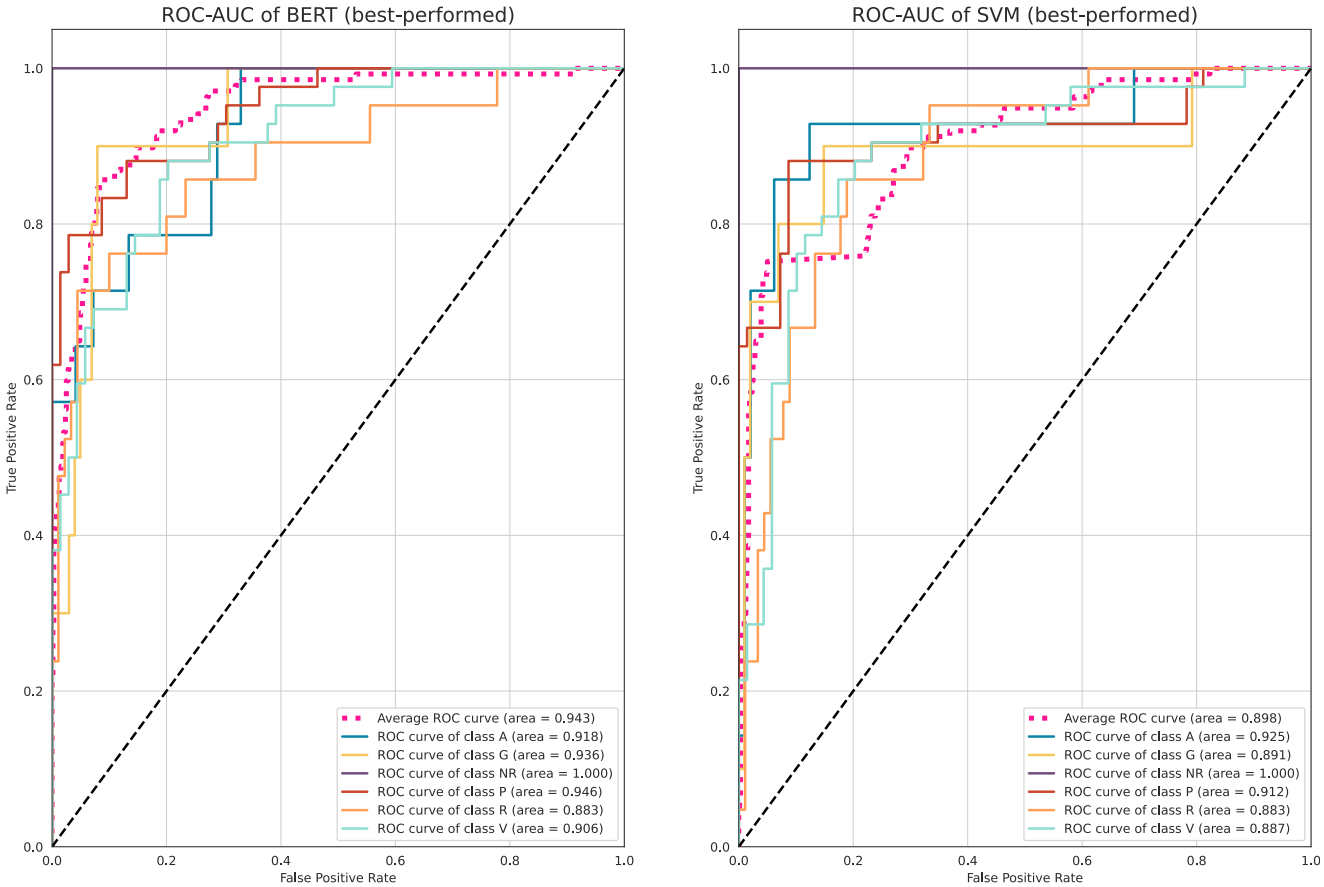


Fig. 4 Comparison of ROC–AUC between BERT and SVM. Notice that the average ROC curve covers higher percentage of area than that from the SVM (0.943 vs. 0.898), indicating that the BERT performed better than SVM

Figure 6 depicts BERT's explanation on the sentence "The heat is traveling up the ruler because the heat is moving from the thumbs onto the ruler. Here the heat of the thumbs is transferring onto both rulers." BERT predicted that the probability of labeling this response into "color heat map visualization (V)" and "haptic perception (P)" classifications is 73% and 58%, respectively. The threshold for multiple labels is 27.2%; thus, this response is coded as the student used visual and perceptual representations as evidence to describe their augmented observations.

Table 2 Results of the multilabel classification

Model	Feature	ROC–AUC	One-error	Coverage error	Ranking loss
*BERT**	*Raw text*	*0.943*	*0.225*	*1.667*	*0.071*
BERT	POS	0.914	0.315	1.829	0.118
BERT	NER	0.927	0.279	1.748	0.079
BERT	POS + NER	0.842	0.369	2.414	0.151
SVM	Raw text	0.896	0.297	1.982	0.115
SVM	*POS*	*0.898*	*0.252*	*1.874*	*0.096*
SVM	NER	0.895	0.279	1.955	0.11
SVM	POS + NER	0.895	0.270	1.901	0.101

*The best-performed model among all. Data in italics stand for the best-performed model using the same algorithm

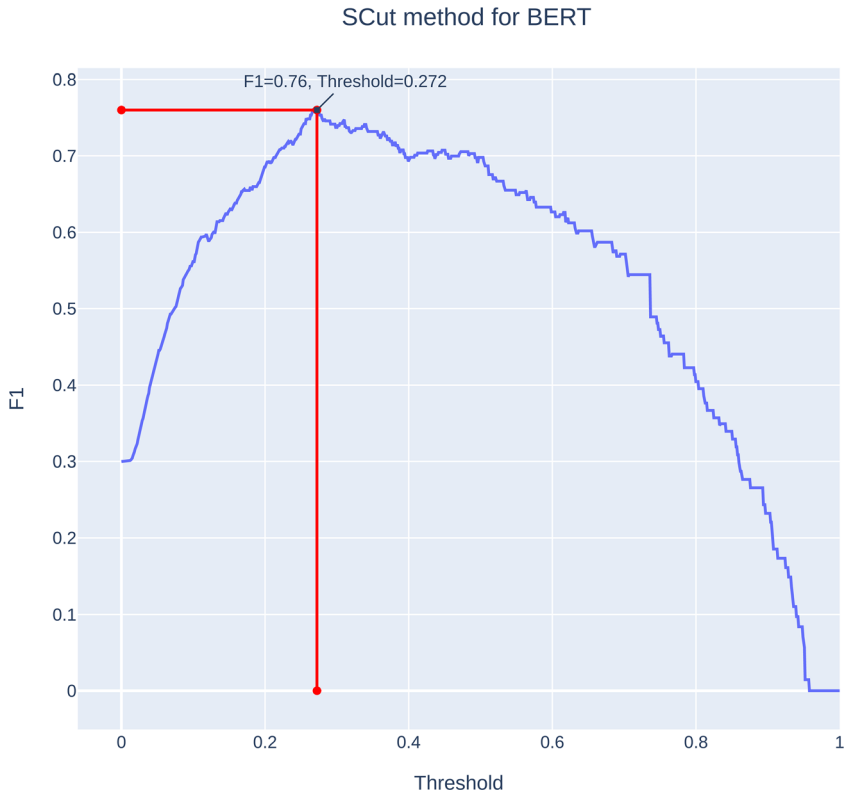


Fig. 5 Searching for the threshold value of BERT using SCut. The threshold result indicated that if the probability of a two or more modes of representational thinking is greater than 0.272, the student is believed to practice multimodal representational thinking

Discussion

This study speaks directly to redefining science assessment from two perspectives: (1) automate the assessment of a complex student construct and (2) apply deep learning in multilabel text classification to ease human labor.

Science educators often infer student's acquisition of representational thinking through the indirect conceptual learning outcome. In our previous study, we examined student's multimodal representational thinking during an AR-assisted laboratory (Sung et al. 2020). The heat conduction lab used sensing technology to augment observation during the POE cycle. We found that students used multiple representations when they responded to the lab report questions. While the findings proved that students indeed applied the representations empowered by the AR technology, the coding process was tedious and labor-intensive. The qualitative data analysis was especially daunting because there were multiple possible codes corresponding to each of the student's responses. Reviewing open-ended responses is innately arduous and costly (Tanana et al. 2016). Even though STEM educators proved that machine learning–based NLP is useful in automated assessment (e.g., Ha et al. 2011; Krithika and

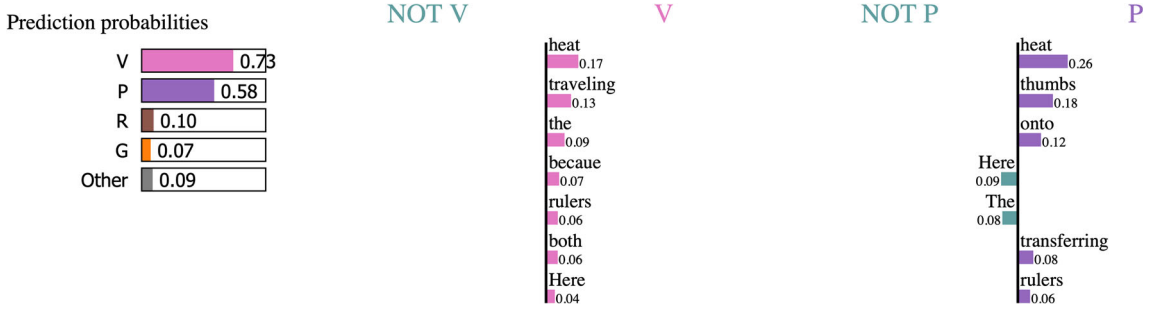


Fig. 6 The visualization of how BERT performs deep learning in assigning multiple labels to the student response

Narayanan 2015; Li et al. 2017; Liu et al. 2016a), very few, if any, reported how they automated the assessment of complex student constructs, such as multimodal representations. In technical terms, most of the automated assessment studies focused on assigning a single score to each analysis unit, while the present study successfully trained the machine to accurately assign multiple codes to each response.

Searching for meaningful themes and valid categories in massive qualitative data requires iterative discussion among the analysts (DeLyser et al. 2013). The machine learning approach is comparable to humans, if not exceeding human interrater reliability (e.g., Ha et al. 2011). It could replicate this time-consuming process with satisfactory accuracy but in a much shorter time. Deep learning (i.e., BERT), which is a subset of the machine learning approach, proves to be more accurate than traditional machine learning methods (i.e., SVM) based on the ROC–AUC scores. The nature of BERT's word embedding approach and the strength of transfer learning architecture might contribute to its satisfactory precision in assessing complex student constructs. BERT considers contextualized meaning of a keyword instead of its standalone definition, and the possibility of leveraging transfer learning with BERT ensures an outstanding base performance on language understanding. Taking the result for Fig. 6 for instance, "heat of the thumb" is an important feature for the "haptic perception (P)" code, and "heat traveling up the ruler" is a critical symbol for "color heat map visualization (V)" code regardless of the duplicative concept "heat" was used in these two response segments. BERT's predicted codes coincide with human coding, and the visualization of logic behind the multiple-label classifications also reassures the validity of the BERT model. BERT's feature engineering mimics the decision-making process behind human coders. The training process for BERT, however, is much more efficient than that in training human coders (Sung et al. 2020). In addition, the deep learning model adopted to automatically assess multiple modes of representational thinking is a milestone. To the best of our knowledge, this paper is the first report of applying multilabel text classification in educational settings. Since the multilabel model was not applied to the testing dataset before training, the fact that the label-prediction results showed fair agreement with human coders indicated BERT is a promising model in decoding complex constructs.

The automated assessment could also speed up the identification of the complex constructs, such as multimodal representational thinking used by students to explain the haptic perception and the artifacts generated in an AR-assisted lab. After specifying important combinations of terms and context-embedding, student's use of the AR features could be automatically captured by the machine. Delegating the assessment of multimodal representational thinking to the machine is advantageous. Practitioners, educational researchers, and AR-app developers would benefit from the information populated by the machine. We will elaborate on how these personnel could take advantage of the automated assessment in an AR-assisted lab below.

Teachers could promptly obtain and use the results generated by the deep learning model to differentiate student's representational thinking mode. BERT would be an effective technique to provide just-in-time feedback for the instructors. The automated analysis of the constructed responses could reveal learners' cognitive engagement and representational thinking. Instructors, relieved from manual grading duty, could now quickly identify disengaged students and take necessary actions for early intervention. Also, teachers could dedicate more time to address a wide spectrum of tasks ranging from monitoring common learning difficulties faced by the group, to remedial pedagogies for individual students. The latter task on tailoring feedback for a particular student is especially beneficial for those who struggle in the lab. These students might be stuck with a single mode of representational thinking, and the lab instructor could redirect these students to diverge thinking accordingly. As noted, heavy reliance on a particular representation tends to impose higher extraneous cognitive demands than diversifying representational input (Mayer and Moreno 2003).

Educational researchers could immediately identify themes to classify student responses and feed the useful information to the AR technology developers. Software developers could take this advice to revise the app and enhance the user's experience accordingly. The machine would be a master "apprentice" in assisting researchers to process text data in large-scale educational research. The machine could perform the task tirelessly with satisfactory results. The massive amount of information processing capacity makes the machine a perfect artificial intelligence (AI) tutor during the hands-on science lab. The researchers and developers could integrate the automated assessment in the "smart" laboratories to provide immediate formative assessment. Adopting machine algorithm as an AI tutor in a hands-on lab is revolutionary; however, machine learning has been widely used in computer-supported collaborative learning, interactive synthetic tutoring system, and massive open online courses that create tremendous amount of behavior data (Crossley et al. 2016; Hew et al. 2018; McNamara et al. 2017; Nakamura et al. 2016). The scientists who study cognitive theory of multimedia learning could be rest assured that the complex representational thinking inferred from the qualitative data analysis is based on the rigorous deep learning result. The machine instantly generates participant's user experience data, so cognitive scientists could decipher whether students' cognitive capacity is overloaded or understimulated.

Conclusions and Future Studies

Our study would contribute to the field of cognitive sciences in multimedia learning, especially from the perspective of how students constructed and manipulated representations offered by AR technology in real-time laboratory sessions. Specifically, studies show that the conjunction of sensory stimuli (i.e., augmented observation features and tactile experience) would enhance learning outcomes (Clark and Jorde 2004) and boost positive attitudes on socioscientific issues with AR-assisted learning (Chang et al. 2018). The AR-assisted laboratory displays the learning of thermal energy concepts that is hard to grasp otherwise. External representations need to be used with caution because unclear multiple representations might impose an extraneous cognitive load during the learning process (Cook 2006). Therefore, explicit tutorials should be included in a "smart" laboratory to coach students how to make connections between different modes of representations and the target concepts. We will offer this type of tutorial in the future to see whether students adopt more transductive and temperature-temporal augmentation in the AR-assisted thermodynamics lab (see Fig. 2b, c).

Since representational thinking is believed to be imperative for scientific modeling and evidence-based reasoning (Gobert and Pallant 2004; Hallström and Schönborn 2019; National Research Council 2008), researchers could investigate how different modes of representational thinking functions during modeling or evidence-based reasoning processes (Sung et al. 2020; Brown et al. 2010). Student's written responses could be quickly quantified based on their quality of reasoning. The complex cognitive processes could be assessed by number and accuracy of evidence provided (e.g., Brown et al. 2010). Interested researchers could also study how students adopt different sources of representational thinking to support their scientific claim. Further investigation could be conducted on how students using more modes of representational thinking in an AR-assisted lab might also perform better in evidence-based reasoning or model-based learning.

Implications and Limitations

The implication for this finding is that the text recognition function of BERT confirms and works hand-in-hand with cognitive science theories. Practitioners could be comforted that even a machine relies heavily on human-defined algorithms and, thus, does not think by itself. We could train the machine to learn based on foundational theories of interest. Also, when we provide clear operational definitions of each mode of representation, machine apprentices, such as BERT, could transform into a master to process big data generated during an AR-assisted lab with precision. Although there are no definite criteria on how much data is needed for deep learning, previous research suggested that traditional machine learning models tended to significantly outperform deep learning when the training sample size is small (Tang et al. 2018). However, this limitation of small training data size is relieved by the adoption of the transfer learning model (Lan et al. 2019). Previously, deep learning models only allowed transfer learning on word embeddings, in which researchers could utilize word feature extraction results trained on a tremendous dataset (Howard and Ruder 2018; Pennington et al. 2014; Peters et al. 2018; Rong 2014). BERT, on the other hand, initiated the breakthroughs on transferring all parameters to end-task such as classification and language generation (Devlin et al. 2018). Recent studies suggested transfer learning models based on BERT have achieved satisfactory results on NLP with limited datasets (Bao and Qiao 2019; Lan et al. 2019). Admittedly, our current dataset is not sufficient enough to evaluate the model effectively; thus, future research will re-evaluate the model with larger datasets.

We did not directly investigate the conceptual change of thermal energy or confirm student's representational thinking with follow-up interviews in this study; however, we believe that teachers and researchers could take advantage of the augmented observation features in an AR-assisted lab with carefully crafted conceptual assessment and interview questions similar to Magana and Balachandran's (2017) article about representational competence. We recommend that instructors present a scenario that imposes cognitive conflict, and then students could apply their newly acquired "super-sense" to reconstruct a more accurate representation of thermal perception. Such cognitive-conflicting moments might become the prime time for a conceptual change (Başer 2006).

The study is based on a one snapshot pre- and post-test research design without a control group, so the preliminary findings might not be generalizable and need to be interpreted with caution. However, we believe that the application of the IR camera in the laboratory provides an unprecedented experience that could augment observations that could not be achieved with the conventional laboratory setup.

Funding This study received funding from the National Science Foundation (1712676, 1626228, 1512868).

Compliance with Ethical Standards

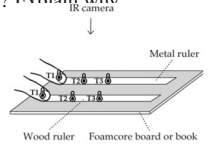
Conflict of Interest The authors declare that they have no conflict of interest.

Ethical Approval No institutional nor author information was revealed in the main text for double blind peer review.

Informed Consent Informed consent was obtained from all individual participants included in the study.

Appendix 1

Table 3 Prompts used in the POE cycle

Prediction	Observation	Explanation
Before you do anything, answer the prediction questions below. 1. When you touch the two rulers, which one will feel cooler? Explain why. 2. What will happen to the temperature patterns of the two rulers after you touch them for 1 minute (as shown in the figure below)? Explain why. 3. What temperature pattern will happen to *your two fingers* right after you move them away from the rulers? Explain why.  	Insert the final graph $T(t)$*: • Which T in the graph are the thermometers on the wood ruler? Which are the ones on the metal ruler? • Write a short description of the thermal energy pattern that the graph is displaying. Insert infrared images of thermal patterns observed: a. Insert the infrared image of the *initial state of two rulers*, and briefly describe the pattern you see. b. Insert the infrared image of the *state of two rulers with thumbs pressing on them for 60 s*, and briefly describe the pattern you observed. c. Insert the infrared image of the *state of two rulers immediately after thumbs move away*, and briefly describe the pattern you observed. d. Insert the infrared image of the *state of thumbs immediately after they move away from rulers*, and briefly describe the pattern you observed.	Use the IR images that you and your partner have taken as evidence to check your prediction. Whether your prediction is correct or not, explain your observation based on the concepts of thermal conduction and thermal conductivity. 1. When you touched the two rulers, which one felt cooler? Explain why. 2. As the image in observation (b) shows, the two rulers' temperatures were different at the places 2 inches away from the thumbs. Explain why. 3. As the image in observation (c) shows, the places two thumbs touched had different temperatures. Explain why. 4. As the image in observation (d) shows, one thumb was cooler than the other after touching two rulers. Explain why.

*$T(t)$ represents the temperature–time graph

Appendix 2

Table 4 Coding scheme of the representations of thermal perceptions for coders

Code	Definition	Sample response
Thermometer reading (R)	• Student's response mentions specific temperature reading. • Student refers to the term "temperature" in their response. • Student compares different temperatures in their response.	• The wooden ruler was 31 degrees and the metal ruler was 29 degrees. The results were around the same. • The rulers are about the same temperature because they were not being touched/they were both at room temperature.
Color heat map visualization (V)	Student's response mentions visual aid or color with reference to the position, location, change of color, spreading, directional, time-related term, quantity-related term, etc.	• The rulers are getting warmer as the thumbs presses down on them. The metal distributes the heat more evenly than the wooden ruler. • The metal one conducted heat away from my hand faster [P, V], meaning it heated up faster • One thumb was cooler than the other thumb because the wooden ruler absorbed much more heat opposed to the metal one.
Temperature–time graph (G)	• Student's response to the prompt mentions about the change of temperature on the graph. • Student responds to the prompt about the graph directly. • Student's response incorporates "temperature" and "rate."	• In the beginning, the temperature increases in a short time period. • The temperature of the metal ruler was rising faster than that of the wooden ruler.
Perception (P)	• Student's response mentions their perception of temperature. • Student's response mentions body parts (e.g., thumb, hand) that includes sense of temperature.	• One thumb is cooler than the other thumb [P] because one ruler allows heat to travel through it better than the other ruler [V].
Abstract (A) conception	Student's response ONLY infers abstract concepts, which cannot be categorized into any other codes.	... will retain your body heat temp but soon return to the temperature of the room reaching thermal equilibrium
Other (O)	Irrelevant responses that do not address the question directly.	• The graph did not come out very clearly, so it is hard to tell which is which. • I do not know.

References

Ainsworth, S. (2008). The educational value of multiple-representations when learning complex scientific concepts. In J. K. Gilbert, M. Reiner, & M. Nakhleh (Eds.), *Visualization: theory and practice in science education* (pp. 191–208). Dordrecht: Springer. https://doi.org/10.1007/978-1-4020-5267-5_9.

Alalga, A., Benabdeslem, K., & Taleb, N. (2016). Soft-constrained Laplacian score for semi-supervised multi-label feature selection. *Knowledge and Information Systems, 47*(1), 75–98.

Allen, L. K., Perret, C., & McNamara, D. S. (2016). Linguistic signatures of cognitive processes during writing. *In Proceedings of the 38th annual meeting of the Cognitive Science Society* (pp. 2483-2488).

Al-Otaibi, R., Flach, P., & Kull, M. (2014). Multi-label classification: a comparative study on threshold selection methods. In *First international workshop on Learning over Multiple Contexts* (LMCE) at ECML-PKDD 2014.

Bailyn, L. (1977). Research as a cognitive process: implications for data analysis. *Quality and Quantity, 11*(2), 97–117. https://doi.org/10.1007/BF00151906.

Bao, X., & Qiao, Q. (2019). Transfer learning from pre-trained BERT for pronoun resolution. In *Proceedings of the first workshop on gender bias in natural language processing* (pp. 82-88).

Başer, M. (2006). Fostering conceptual change by cognitive conflict based instruction on students' understanding of heat and temperature concepts. *Eurasia Journal of Mathematics, Science and Technology Education, 2*(2), 96–114. https://doi.org/10.12973/ejmste/75458.

Beggrow, E. P., Ha, M., Nehm, R. H., Pearl, D. K., & Boone, W. J. (2014). Assessing scientific practices using machine-learning methods: how closely do they match clinical interview performance? *Journal of Science Education and Technology, 23*, 160–182. https://doi.org/10.1007/s10956-013-9461-9.

Blank, G. (2004). Teaching qualitative data analysis to graduate students. *Social Science Computer Review, 22*(2), 187–196. https://doi.org/10.1177/0894439303262559.

Brown, N., Furtak, E., Timms, M., Nagashima, S., & Wilson, M. (2010). The evidence-based reasoning framework: assessing scientific reasoning. *Educational Assessment, 15*(3-4), 123–141. https://doi.org/10.1080/10627197.2010.530551.

Carbó, P. A., Andrea Vázquez Ahumada, M., Caballero, A. D., & Lezama Argüelles, G. A. (2016). "How do I do discourse analysis?" Teaching discourse analysis to novice researchers through a study of intimate partner gender violence among migrant women. *Qualitative Social Work, 15*(3), 363–379. https://doi.org/10.1177/1473325015617233.

Chang, H.-Y., Hsu, Y.-S., Wu, H.-K., & Tsai, C.-C. (2018). Students' development of socio-scientific reasoning in a mobile augmented reality learning environment. *International Journal of Science Education, 40*(12), 1410–1431. https://doi.org/10.1080/09500693.2018.1480075.

Clark, D., & Jorde, D. (2004). Helping students revise disruptive experientially supported ideas about thermodynamics: computer visualizations and tactile models. *Journal of Research in Science Teaching, 41*(1), 1–23.

Clough, E. E., & Driver, R. (1985). Secondary students' conceptions of the conduction of heat: bringing together scientific and personal views. *Physics Education, 20*(4), 176–182.

Cook, M. P. (2006). Visual representations in science education: the influence of prior knowledge and cognitive load theory on instructional design principles. *Science Education, 90*(6), 1073–1091. https://doi.org/10.1002/sce.20164.

Crossley, S., Paquette, L., Dascalu, M., McNamara, D. S., & Baker, R. S. (2016, April). Combining click-stream data with NLP tools to better understand MOOC completion. In *Proceedings of the sixth international conference on learning analytics & knowledge* (pp. 6-14). ACM.

DeLyser, D., Potter, A. E., Chaney, J., Crider, S., Debnam, I., Hanks, G., Hotard, C. D., Modlin, E. A., Pfeiffer, M., & Seemann, J. (2013). Teaching qualitative research: experiential learning in group-based interviews and coding assignments. *Journal of Geography, 112*(1), 18–28. https://doi.org/10.1080/00221341.2012.674546.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv*:1810.04805.

Donnelly, D. F., Vitale, J. M., & Linn, M. C. (2015). Automated guidance for thermodynamics essays: critiquing versus revisiting. *Journal of Science Education and Technology, 24*(6), 861–874.

Dunleavy, M., Dede, C., & Mitchell, R. (2009). Affordances and limitations of immersive participatory augmented reality simulations for teaching and learning. *Journal of Science Education and Technology, 18*(1), 7–22. https://doi.org/10.1007/s10956-008-9119-1.

Ebenezer, J., Chacko, S., Kaya, O. N., Koya, S. K., & Ebenezer, D. L. (2010). The effects of common knowledge construction model sequence of lessons on science achievement and relational conceptual change. *Journal of Research in Science Teaching, 47*(1), 25–46.

Este, D., Sieppert, J., & Barsky, A. (1998). Teaching and learning qualitative research with and without qualitative data analysis software. *Journal of Research on Computing in Education, 31*(2), 138–154. https://doi.org/10.1080/08886504.1998.10782247.

Fan, J., Upadhye, S., & Worster, A. (2006). Understanding receiver operating characteristic (ROC) curves. *Canadian Journal of Emergency Medicine, 8*(1), 19–20.

Fedus, W., Goodfellow, I., & Dai, A. M. (2018). MaskGAN: better text generation via filling in the_. In *Proceedings of the sixth International Conference on Learning Representations* (ICLR).

Feng, Y., Jones, J., Chen, Z., & Fang, C. (2018). An empirical study on software failure classification with multi-label and problem-transformation techniques. In *2018 IEEE 11th International Conference on Software Testing, Verification and Validation (ICST)* (pp. 320-330). IEEE.

Garzón, J., & Acevedo, J. (2019). Meta-analysis of the impact of augmented reality on students' learning gains. *Educational Research Review, 27*(1), 244–260. Retrieved October 31, 2019 from https://www.learntechlib.org/p/209849/.

Gobert, J. D., & Pallant, A. (2004). Fostering students' epistemologies of models via authentic model-based tasks. *Journal of Science Education and Technology, 13*(1), 7–22. https://doi.org/10.1023/B:JOST.0000019635.70068.6f.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Grolemund, G., & Wickham, H. (2014). A cognitive interpretation of data analysis. *International Statistical Review, 82*(2), 184–204. https://doi.org/10.1111/insr.12028.

Ha, M., Nehm, R. H., Urban-Lurain, M., & Merrill, J. E. (2011). Applying computerized-scoring models of written biological explanations across courses and colleges: prospects and limitations. *CBE Life Sciences Education, 10*(4), 379–393. https://doi.org/10.1187/cbe.11-08-0081.

Hallström, J., & Schönborn, K. J. (2019). Models and modelling for authentic STEM education: reinforcing the argument. *International Journal of STEM Education, 6*(1). https://doi.org/10.1186/s40594-019-0178-z.

Haudek, K. C., Prevost, L. B., Moscarella, R. A., Merrill, J., & Urban-Lurain, M. (2012). What are they thinking? Automated analysis of student writing about acid–base chemistry in introductory biology. *CBE Life Sciences Education, 11*(3), 283–293. https://doi.org/10.1187/cbe.11-08-0084.

Hew, K. F., Qiao, C., & Tang, Y. (2018). Understanding student engagement in large-scale open online courses: a machine learning

facilitated analysis of student's reflections in 18 highly rated MOOCs. *The International Review of Research in Open and Distance Learning, 19*(3), 69–93. Retrieved March 6, 2020 from http://ezproxy.lib.utexas.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=EJ1185116&site=ehost-live.

Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. *arXiv preprint arXiv*:1801.06146.

Jitendra, A. K., Nelson, G., Pulles, S. M., Kiss, A. J., & Houseworth, J. (2016). Is mathematical representation of problems an evidence-based strategy for students with mathematics difficulties? *Exceptional Children, 83*(1), 8–25. https://doi.org/10.1177/0014402915625062.

Kastrati, Z., Imran, A. S., & Kurti, A. (2019). Transfer learning to timed text based video classification using CNN. In *Proceedings of the 9th international conference on web intelligence, mining and semantics* (pp. 1-9).

Kovanović, V., Joksimović, S., Gašević, D., & Hatala, M. (2014). Automated content analysis of online discussion transcripts. In *Proceedings of the workshops at the LAK 2014 conference co-located with 4th international conference on Learning Analytics and Knowledge* (LAK 2014).

Krithika, R., & Narayanan, J. (2015). Learning to grade short answers using machine learning techniques. *Proceedings of the Third International Symposium on Women in Computing and Informatics*, 262–271. https://doi.org/10.1145/2791405.2791508

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). ALBERT: a lite BERT for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444.

Lewis, E. L., & Linn, M. C. (1994). Heat energy and temperature concepts of adolescents, adults, and experts: implications for curricular improvements. *Journal of Research in Science Teaching, 31*(6), 657–677.

Li, H., Gobert, J. D., & Dickler, R. (2017). Automated assessment for scientific explanations in on-line science inquiry. EDM.

Li, N., Liu, S., Liu, Y., Zhao, S., Liu, M., & Zhou, M. (2018). Close to human quality TTS with transformer. *arXiv preprint arXiv*: 1809.08895.

Lilleberg, J., Zhu, Y., & Zhang, Y. (2015, July). Support vector machines and word2vec for text classification with semantic features. In *2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC)* (pp. 136-140). IEEE.

Liu, O. L., Rios, J. A., Heilman, M., Gerard, L., & Linn, M. C. (2016a). Validation of automated scoring of science assessments. *Journal of Research in Science Teaching, 53*(2), 215–233. https://doi.org/10.1002/tea.21299.

Liu, Y., Sun, C., Lin, L., & Wang, X. (2016b). Learning natural language inference using bidirectional LSTM model and inner-attention. arXiv preprint arXiv:1605.09090.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Advances in neural information processing systems* (pp. 4765-4774).

Luo, W., & Litman, D. (2016, March). Determining the quality of a student reflective response. In The twenty-ninth international FLAIRS Conference.

Magana, A. J., & Balachandran, S. (2017). Students' development of representational competence through the sense of touch. *Journal of Science Education and Technology, 26*(3), 332–346. https://doi.org/10.1007/s10956-016-9682-9.

Mayer, R. E. (2014). Cognitive theory of multimedia learning. In *The Cambridge handbook of multimedia learning* (2nd ed., pp. 43–71). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139547369.005.

Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist, 38*(1), 43–52. https://doi.org/10.1207/S15326985EP3801_6.

McNamara, D. S., Allen, L., Crossley, S., Dascalu, M., & Perret, C. A. (2017). Natural language processing and learning analytics. Handbook of learning analytics, 93–104.

Mehmood, A., On, B. W., Lee, I., & Choi, G. (2017). Prognosis essay scoring and article relevancy using multi-text features and machine learning. *Symmetry, 9*(1), 11.

Mitchell, T. M. (1997). *Machine learning*. Burr Ridge, IL: McGraw Hill. Retrieved March 6, 2020 from https://www.cs.ubbcluj.ro/~gabis/ml/mlbooks/McGrawHill%20-%20Machine%20Learning%20-Tom%20Mitchell.pdf.

Moore, T. J., Brophy, S. P., Tank, K. M., Lopez, R. D., Johnston, A. C., Hynes, M. M., & Gajdzik, E. (2020). Multiple representations in computational thinking tasks: a clinical study of second-grade students. *Journal of Science Education and Technology, 29*(1), 19–34. https://doi.org/10.1007/s10956-020-09812-0.

Moreno, R. (2006). Does the modality principle hold for different media? A test of the method-affects-learning hypothesis. *Journal of Computer Assisted Learning, 22*(3), 149–158. https://doi.org/10.1111/j.1365-2729.2006.00170.x.

Mou, L., Meng, Z., Yan, R., Li, G., Xu, Y., Zhang, L., & Jin, Z. (2016). How transferable are neural networks in nlp applications?. *arXiv preprint arXiv*:1603.06111.

Nakamura, C. M., Murphy, S. K., Christel, M. G., Stevens, S. M., & Zollman, D. A. (2016). Automated analysis of short responses in an interactive synthetic tutoring system for introductory physics. *Physical Review Physics Education Research, 12*(1), 010122. https://doi.org/10.1103/PhysRevPhysEducRes.12.010122.

Nanaware, R. (2018). Chatbot for education system. *International Journal of Emerging Technology and Computer Science, 3*(2), 43–48.

Namdar, B., & Shen, J. (2015). Modeling-oriented assessment in k-12 science education: A synthesis of research from 1980 to 2013 and new directions. *International Journal of Science Education, 37(7)*, 993–1023. https://doi.org/10.1080/09500693.2015.1012185.

Nareshpalsingh, M. J., & Modi, N. H. (2017). Multi-label classification methods: a comparative study. *International Research Journal of Engineering and Technology (IRJET), 4*(12), 263–270.

National Research Council. (2008). *Ready, set, SCIENCE!: putting research to work in K-8 science classrooms* (p. 10.17226/11882). Washington: The National Academies Press.

Nayak, A., & Natarajan, D. (2016). Comparative study of naive Bayes, support vector machine and random forest classifiers in sentiment analysis of twitter feeds. *International Journal of Advance Studies in Computer Science and Engineering (IJASCSE), 5*(1), 16.

Nehm, R. H., Ha, M., & Mayfield, E. (2012). Transforming biology assessment with machine learning: automated scoring of written evolutionary explanations. *Journal of Science Education and Technology, 21*(1), 183–196. https://doi.org/10.1007/s10956-011-9300-9.

Next Generation Science Assessment Task Collections. (2019). *Explore and use our classroom-ready assessment tasks.* Retrieved July 21, 2019 from http://nextgenscienceassessment.org/.

Next Generation Science Standards (NGSS) Lead States. (2013). *Next generation science standards: for states, by states*. Washington: The National Academies Press.

Pathare, S. R., & Pradhan, H. C. (2010). Students' misconceptions about heat transfer mechanisms and elementary kinetic theory. *Physics Education, 45*(6), 629–634.

Patton, M. Q. (2005). Qualitative research. In *Encyclopedia of statistics in behavioral science*. Retrieved from https://doi.org/10.1002/0470013192.bsa514

Pennacchiotti, M., & Popescu, A. M. (2011). A machine learning approach to twitter user classification. In *Fifth international AAAI conference on weblogs and social media*.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: global vectors for word representation. In Proceedings of the 2014 conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1532-1543).

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv*:1802.05365.

Pla, F., & Hurtado, L. F. (2014). Sentiment analysis in Twitter for Spanish. In *In International conference on applications of natural language to data bases/information systems* (pp. 208–213). Cham: Springer.

Pranckevičius, T., & Marcinkevičius, V. (2017). Comparison of naive Bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification. *Baltic Journal of Modern Computing, 5*(2), 221.

Reiner, M., Slotta, J. D., Chi, M. T. H., & Resnick, L. B. (2000). Naive physics reasoning: a commitment to substance-based conceptions. *Cognition and Instruction, 18*(1), 1–3.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135–1144).

Rong, X. (2014). word2vec parameter learning explained. *arXiv preprint arXiv*:1411.2738.

Ruder, S., Peters, M. E., Swayamdipta, S., & Wolf, T. (2019). Transfer learning in natural language processing. In *Proceedings of the 2019 conference of the North American Chapter of the Association for Computational Linguistics: tutorials* (pp. 15-18).

Samuelsson, C. R., Elmgren, M., Xie, C., & Haglund, J. (2019). Going through a phase: Infrared cameras in a teaching sequence on evaporation and condensation. *American Journal of Physics*, *87*(7), 577–582. https://doi.org/10.1119/1.5110665.

Sarker, A., Mollá, D., & Paris, C. (2013). An approach for automatic multi-label classification of medical sentences. In *Proceedings of the 4th international Louhi Workshop on Health Document Text Mining and Information Analysis*. Sydney, NSW, Australia.

Schenk, N., Chiarcos, C., Donandt, K., Rönnqvist, S., Stepanov, E., & Riccardi, G. (2016). Do we really need all those rich linguistic features? A neural network-based approach to implicit sense labeling. In *Proceedings of the CoNLL-16 shared task* (pp. 41–49).

Shen, J., & Confrey, J. (2007). From conceptual change to transformative modeling: A case study of an elementary teacher in learning astronomy. *Science Education*, *91*(6), 948–966. https://doi.org/10.1002/sce.20224.

SpolaôR, N., Cherman, E. A., Monard, M. C., & Lee, H. D. (2013). A comparison of multi-label feature selection methods using the problem transformation approach. *Electronic Notes in Theoretical Computer Science, 292*, 135–151.

Sun, Z., Guo, Z., Liu, C., Wang, X., Liu, J., & Liu, S. (2017). Fast extended one-versus-rest multi-label support vector machine using approximate extreme points. *IEEE Access, 5*, 8526–8535.

Sung, S., Huang, X., Shen, J., Wang, C., Xie, C., Zeng, Y. & Chen, G. (2020, Apr 17 - 21) *Augmented Visual Perception: Interpreting Thermal Sensation with Innovative Technology* [Poster Session]. American Educational Research Association (AERA) Annual Meeting San Francisco, CA http://tinyurl.com/tdgnfwm (Conference Canceled)

Tanana, M., Hallgren, K. A., Imel, Z. E., Atkins, D. C., & Srikumar, V. (2016). A comparison of natural language processing methods for automated coding of motivational interviewing. *Journal of Substance Abuse Treatment, 65*, 43–50. https://doi.org/10.1016/j.jsat.2016.01.006.

Tang, A., Tam, R., Cadrin-Chênevert, A., Guest, W., Chong, J., Barfett, J., et al. (2018). Canadian Association of Radiologists white paper on artificial intelligence in radiology. *Can Assoc Radiol J, 69*(2), 120–135.

Tsoumakas, G., & Katakis, I. (2007). Multi-label classification: an overview. *International Journal of Data Warehousing and Mining, 3*(3), 1–13.

Tsoumakas, G., Katakis, I., & Vlahavas, I. (2009). Mining multi-label data. In *In Data mining and knowledge discovery handbook* (pp. 667–685). Boston: Springer.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998–6008).

Walsh, Y., Magana, A. J., & Feng, S. (2020). Investigating students' explanations about friction concepts after interacting with a visuohaptic simulation with two different sequenced approaches. *Journal of Science Education and Technology, 29*(4), 443–458. https://doi.org/10.1007/s10956-020-09829-5.

Wiser, M., & Kipman, D. (1988). *The differentiation of heat and temperature: an evaluation of the effect of microcomputer models on students' misconceptions*. In *Tech Report* (pp. 88–20). Cambridge: Educational Technology Center, Harvard University.

Wu, H.-K., & Puntambekar, S. (2012). Pedagogical affordances of multiple external representations in scientific processes. *Journal of Science Education and Technology, 21*(6), 754–767. https://doi.org/10.1007/s10956-011-9363-7.

Wu, X. Z., & Zhou, Z. H. (2017). A unified view of multi-label performance measures. In *Proceedings of the 34th international conference on machine learning-volume 70* (pp. 3780-3788). JMLR. Org.

Wu, H.-K., Lee, S. W.-Y., Chang, H.-Y., & Liang, J.-C. (2013). Current status, opportunities and challenges of augmented reality in education. *Computers in Education, 62*, 41–49. https://doi.org/10.1016/j.compedu.2012.10.024.

Xie, C. (2011). Visualizing Chemistry with Infrared Imaging. *Journal of Chemical Education*, *88*(7), 881–885. https://doi.org/10.1021/ed1009656.

Xing, W., & Gao, F. (2018). Exploring the relationship between online discourse and commitment in Twitter professional learning communities. *Computers in Education, 126*, 388–398.

Xu, J. (2011). An extended one-versus-rest support vector machine for multi-label classification. *Neurocomputing, 74*(17), 3114–3124.

Yeh, C. K., Wu, W. C., Ko, W. J., & Wang, Y. C. F. (2017). Learning deep latent space for multi-label classification. In *Thirty-first AAAI conference on artificial intelligence*.

Zehner, F., Sälzer, C., & Goldhammer, F. (2016). Automatic coding of short text responses via clustering in educational assessment. *Educational and Psychological Measurement, 76*(2), 280–303.

Zhang, M. L., & Zhou, Z. H. (2013). A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering, 26*(8), 1819–1837.

Zhang, M. L., Li, Y. K., Liu, X. Y., & Geng, X. (2018a). Binary relevance for multi-label learning: an overview. *Frontiers of Computer Science, 12*(2), 191–202.

Zhang, B., Xiong, D., & Su, J. (2018b). Accelerating neural transformer via an average attention network. *arXiv preprint arXiv*:1805.00631.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.