Check for updates

Received: 22 November 2021 | Accepted: 10 June 2022

DOI: 10.1111/bjet.13253

ORIGINAL ARTICLE

British Journal of Educational Technology BERA

An empirical analysis of high school students' practices of modelling with unstructured data

Shiyan Jiang[1] | **Amato Nocera**[1] | **Cansu Tatar**[1] | **Michael Miller Yoder**[2] | **Jie Chao**[3] | **Kenia Wiedemann**[3] | **William Finzer**[3] | **Carolyn P. Rosé**[2]

[1]Department of Teacher Education and Learning Sciences, North Carolina State University, Raleigh, North Carolina, USA

[2]School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA

[3]Concord Consortium, Concord, Massachusetts, USA

Correspondence
Shiyan Jiang, Department of Teacher Education and Learning Sciences, North Carolina State University, POE Hall, 310-G, Raleigh, NC, USA.
Email: sjiang24@ncsu.edu

Funding information
National Science Foundation

Abstract

To date, many AI initiatives (eg, AI4K12, CS for All) developed standards and frameworks as guidance for educators to create accessible and engaging Artificial Intelligence (AI) learning experiences for K-12 students. These efforts revealed a significant need to prepare youth to gain a fundamental understanding of how intelligence is created, applied, and its potential to perpetuate bias and unfairness. This study contributes to the growing interest in K-12 AI education by examining student learning of modelling real-world text data. Four students from an Advanced Placement computer science classroom at a public high school participated in this study. Our qualitative analysis reveals that the students developed nuanced and in-depth understandings of how text classification models—a type of AI application—are trained. Specifically, we found that in modelling texts, students: (1) drew on their social experiences and cultural knowledge to create predictive features, (2) engineered predictive features to address model errors, (3) described model learning patterns from training data and (4) reasoned about noisy features when comparing models. This study contributes to an initial understanding of student learning of modelling unstructured data and offers implications for scaffolding in-depth reasoning about model decision making.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.
© 2022 The Authors. *British Journal of Educational Technology* published by John Wiley & Sons Ltd on behalf of British Educational Research Association.

14678535, 2022, 5, Downloaded from https://bera-journals.onlinelibrary.wiley.com/doi/10.1111/bjet.13253, Wiley Online Library on [30/05/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

KEYWORDS

feature engineering, machine learning, model decision making, unstructured data

Practitioner notes

What is already known about this topic

- Scholarly attention has turned to examining Artificial Intelligence (AI) literacy in K-12 to help students understand the working mechanism of AI technologies and critically evaluate automated decisions made by computer models.
- While efforts have been made to engage students in understanding AI through building machine learning models with data, few of them go in-depth into teaching and learning of feature engineering, a critical concept in modelling data.
- There is a need for research to examine students' data modelling processes, particularly in the little-researched realm of unstructured data.

What this paper adds

- Results show that students developed nuanced understandings of models learning patterns in data for automated decision making.
- Results demonstrate that students drew on prior experience and knowledge in creating features from unstructured data in the learning task of building text classification models.
- Students needed support in performing feature engineering practices, reasoning about noisy features and exploring features in rich social contexts that the data set is situated in.

Implications for practice and/or policy

- It is important for schools to provide hands-on model building experiences for students to understand and evaluate automated decisions from AI technologies.
- Students should be empowered to draw on their cultural and social backgrounds as they create models and evaluate data sources.
- To extend this work, educators should consider opportunities to integrate AI learning in other disciplinary subjects (ie, outside of computer science classes).

INTRODUCTION

The ability to leverage existing data of all forms is increasingly recognized as a skill that is needed in virtually every discipline (Wise, 2020). As a result, efforts to incorporate data literacy and analytics throughout K-16 education are on the rise (Jiang & Kahn, 2020; Lee & Wilkerson, 2018; Philip & Rubel, 2019). In this paper, we argue that a neglected area of data literacy is dealing with unstructured data as well as the restructuring of data to prepare for effective modelling. Data are frequently treated as given, as though it is provided in a neat, tabular form, and the main activity of modelling is the application of mathematical paradigms. For this reason, efforts to increase data literacy and skills related to analytics focus on math and are housed primarily in K-12 math standards (Enyedy & Mukhopadhyay, 2007). However, the reality is that the modeller has a critical role to play in the structuring and restructuring of data to reveal meaningful patterns such that modelling paradigms are able to detect those patterns. In the language of the discipline of machine learning, a subfield of Artificial Intelligence (AI), it is up to the modeller to work through data representation and make patterns learnable (AI4K12, 2019).

14678535, 2022, 5, Downloaded from https://bera-journals.onlinelibrary.wiley.com/doi/10.1111/bjet.13253, Wiley Online Library on [30/05/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

The value of unstructured data like text, image, audio and video is increasingly recognized by scholars and the public (Tatar et al., 2021; Witten et al., 2016). The computational advantage of machine learning paradigms is their ability to learn representations from unstructured data to solve practical problems. However, it is widely acknowledged that a limitation of these paradigms is that the models are frequently opaque. To casual observers, the process of modelling data is masked by machine learning technologies, leading many to view AI as a kind of magic (Long & Magerko, 2020). When it comes to modelling for the purpose of understanding data, these new paradigms fall short. This paper adds to our knowledge of how students learn with data and about data, particularly in the little-researched realm of unstructured data.

The data source this paper focuses on is text from real-world data sets. When models are used for assigning categories to texts, they are called text classification models. Transformation of text into structured data for analysis and modelling is critical in many areas of inquiry such as medicine, education, forensics, and national security (Chowdhury, 2003). Modellers use a variety of techniques such as discretization, reduction and quantification to transform unstructured data into structured data. Each transformation reflects modellers' assumptions about what is important about the data for solving the problem at hand. The transformed data inevitably carry the modellers' assumptions and biases for the purpose of solving the problem. Without insight into the process of discretization, reduction and quantification, students lack the ability to appropriately question assumptions behind the tabular data they are provided with. They are therefore naïve to the extent to which assumptions bias the conclusions that are then drawn from the tabular data. Although there will no doubt be limits to the level of understanding and insight that can be achieved with high school students, the goal is to plant intellectual seeds that can be nurtured and further developed as they move into higher levels of education. This study aims to fill the gap in current research on understanding student learning with unstructured data from real-world data sets. Specifically, the driving research question is as follows: *How do high school students engage in modelling unstructured data in the context of building machine learning models with texts?*

THEORETICAL FRAMEWORK

This study builds on social-cultural theories of learning (Vygotsky, 1978). In this view, data modelling is understood as a culturally mediated process (Wertsch, 1998). In data modelling, one needs to create features from raw, often unstructured, data so that these features can be used for building models (Petrosino, 2016; Rosé, 2018). By approaching this activity as mediated action, we conceptualize data modelling as a culturally significant meaning-making activity (Wertsch, 1998). In this study, students engage in building models with real-world data sets. This theoretical perspective towards data modelling and interaction with real-world data sets and computational technologies helps reveal social aspects of the learning process (Greeno & Engestrom, 2014). It focuses on learner engagement with the technology, including the constraints and affordances of these tools in terms of the kinds of data modelling activity, interaction and learning they can support (Greeno, 1994).

Real-world data sets are compelling resources for supporting the development of students' capabilities in data modelling, offering students the opportunity to use data that is relatable and connected with their lived experience (Hammerman & Rubin, 2004). In our study context, students create features from raw data by drawing on their prior knowledge and personal experiences. The theoretical underpinning of making connections to prior knowledge and cultural backgrounds draws on the idea of sense-making through interactions in complex systems (Zimmerman et al., 2010). In this framework, meaning-making occurs through talk, gesture and engagement with the learning environment and social tools. Students draw on cultural frameworks they learn growing up—along with other ways of knowing—while making meaning from content, especially content involving new or messy information.

RELATED RESEARCH

AI has gained rapid popularity and become a hot topic of debate in the media, policy circles and the academy. Many researchers and educators have started to investigate ways of supporting students' AI understandings through designing activities that involve machine learning and classification (Van Brummelen, 2019; Zhou et al., 2020; Zimmermann-Niefield et al., 2019). When designing such activities and investigating student learning in these activities, the field has paid close attention to selecting appropriate data sets, creating meaningful classification tasks, and revealing the internal process of machine learning algorithms and model decision making.

Data set selection plays a critical role in designing engaging machine learning activities for students. Particularly, unstructured data sets, such as image, text, audio and video, provide rich learning opportunities for students to explore and analyse data to understand or develop machine learning models (Lin et al., 2020; Sakulkueakulsuk et al., 2018; Tang et al., 2019). For this reason, researchers purposefully selected data sets that students might be familiar with or interested in (Shamir & Levin, 2022). For instance, Mobasher et al. (2019) organized a summer academy in which high school students explored a variety of data sets, including bicycle sharing system, human activity recognition, and Spotify song feature data sets. The researchers found that students had productive discussions with these data sets as they provided familiar contexts for students to apply machine learning techniques. As another example, Vartiainen et al. (2020) found that using the body to generate image data sets for classifying emotional responses helped children to reason about the difference between human and model decision making. While these studies demonstrate that unstructured data are essential for machine learning activities, students need more support to learn how to turn unstructured data into recognizable structured data for classification.

In designing machine learning and classification activities, researchers have focused on engaging students in reasoning about automated decisions made by AI technologies by having students build machine learning models (Ho & Scadding, 2019; Lee & Moon, 2020; Marques et al., 2020). In Biehler and Fleischer's study (Biehler & Fleischer, 2021), students created decision trees manually with a selected variable (eg, a decision tree predicting whether a person played online games frequently based on gender information). This study fills the gap of teaching the concept of a decision tree, which is an algorithm suggested by IDSSP for introducing data science in schools (IDSSP Curriculum Team, 2019). They aimed to make the supervised machine learning algorithm of decision trees visible to students by engaging them in manually constructing trees using provided variables (eg, gender information, whether a person played online games frequently and whether a person used Instagram frequently). In another example, Tang et al. (2019) developed a curriculum to teach core machine learning concepts with image classification tasks for high school students. In this study, students took photos of different objects (eg, water bottle, hair tie, and uniform) and trained machine learning models to classify these photos. In addition, students could edit layer parameters to build neural network models. The authors argued that it is important to give students the space to edit models. Collectively, we can see that researchers made algorithms transparent to students and opened the black box of machine learning algorithms by engaging students in creating models.

Although many effective AI curricula have been developed for K-12 (Estevez et al., 2019; Sabuncuoglu, 2020; Williams et al., 2021), only a few of them go in-depth into teaching and learning of feature engineering, a critical concept in machine learning. In one of the few studies, Sakulkueakulsuk et al. (2018) created game-based classification tasks to help middle

14678535, 2022, 5, Downloaded from https://bera-journals.onlinelibrary.wiley.com/doi/10.1111/bjet.13253, Wiley Online Library on [30/05/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

school students to learn AI concepts. In this study, students were guided to identify specific features (eg, softness, colour, and texture) to develop machine learning models for classifying the grade of mangoes. Although students created predictive features that led to high accuracy predictions, this study provided a limited understanding of how students discovered features and students' processes of selecting certain features to build machine learning models. More efforts should be devoted to examining the processes of how students turn unstructured data into meaningful features, which is an important part of this study.

Building from the above theoretical perspectives and literature on K-12 AI education, we examine how students' analysing real-world data sets helps mediate their data modelling activities and how students make sense of the data interfaces to create data models. Specifically, we examine students' step-by-step interaction with data-related technologies to understand their processes moving from constructing features from unstructured data sets (in this study, texts) to building machine learning models.

METHODOLOGY

This study is part of a three-year research and development project called Narrative Modelling with StoryQ. The project aimed to advance students' understanding of machine learning with unstructured data, with a focus on text classification. This paper presents a detailed, descriptive analysis of high school youth engaging in modelling texts using real-world data sets.

Site and participants

The study took place in an Advanced Placement (AP) computer science classroom at a public high school in the Northeastern United States. The composition of the school population was 51% Latinx, 41% White, 4% Black and 4% other. This course was an introductory college-level computer science course with a focus on understanding basic programming concepts and familiarizing students with advanced technologies such as AI.

We use pseudonyms for all the participants in this study. The AP class was taught by Mr. Smith, a male teacher in his 60s. The teacher had previously taught mathematics for 5 years and technology and computing classes for 7 years. He had no experience in artificial intelligence, machine learning, and text mining (ie, building machine learning models with texts) prior to his participation in the study. He received a total of 10 hours of one-on-one training with the research team before implementing this project. Four students, Albert (Latino, 11th grade), Emily (White female, 10th grade), Eric (Latino, 11th grade), and Maya (Asian female, 11th grade), enrolled in the class. Alberto and Eric worked as a pair and Emily and Maya worked as another pair when building models to classify texts. They were confident in learning new technologies and interested in pursuing STEM careers in technology-related fields. None of them had taken any AI classes and they had limited experience discussing AI with others. But they were interested in learning how AI works, learns, is being created and how it might affect their lives.

The AI unit

The class participated in a 3-week AI unit. The implementation of our unit lasted 14 class sessions (one session per day and each session was 1 hour) and was conducted remotely via Zoom (ie, a video conferencing tool) due to COVID-19. Each day, researchers met with the teacher before the class to discuss the lesson agenda. At the beginning of each class

14678535, 2022, 5, Downloaded from https://bera-journals.onlinelibrary.wiley.com/doi/10.1111/bjet.13253, Wiley Online Library on [30/05/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

14678535, 2022, 5, Downloaded from https://bera-journals.onlinelibrary.wiley.com/doi/10.1111/bjet.13253, Wiley Online Library on [30/05/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

session, in the Zoom main room, the teacher informed the students about the lesson plan and then the researchers presented the AI curriculum activities. Afterwards, the students completed the activities as pairs in Zoom breakout rooms and discussed key concepts as a whole class in the main room.

The learning activities introduced machine learning topics through engaging students in hands-on experiences of building data models with text data. Specifically, they consisted of four parts:

- *Introduction to AI*: The research team introduced the concept of machine learning, a subfield of AI, and different types of data (eg, text data) that machine learning models use.
- *Human classifying texts*: Students read restaurant reviews and classified them as either positive or negative, later discussing how they made their decisions.
- *Computers classifying texts*: Students created models with restaurant reviews as text data in StoryQ to classify positive and negative reviews.
- *Leveraging human insights in developing computer models to classify texts*: Students conducted error analysis to improve their models. They were prompted to compare human and computer decision making.

Overall, these activities were designed to help students to reason about human and computer decision making and understand the role and responsibility of humans in creating AI technologies.

Technology: StoryQ

To make machine learning concepts and practices engaging and accessible, we used strategies including interactive visualizations and dynamically linked representations (eg, Sorva et al., 2013). To implement these strategies, we developed StoryQ, a web-based text mining and narrative modelling platform, to support the entire process of the text mining practice in a visual, interactive fashion with dynamically linked representations.

Our curriculum allows users to explore and iteratively develop text classification models in StoryQ. The model used logistic regression over unigram features (ie, single words) as the classification model, which allows predicting restaurant reviews from Yelp into discrete labels (ie, positive and negative) by learning the relationship from a given set of reviews with human-labelled ratings. We chose this model as it is easy to interpret for novices and very efficient to train (Bapat et al., 2018). For instance, Figure 1 shows actions users can take to explore results from text classification models. The classification task is to predict whether Yelp reviews are negative or positive. Users can create a distribution graph (Figure 1b) to interpret reviews in the data set (Figure 1a). The data set has 500 reviews. When clicking a dot in the graph, the row of the selected point will be highlighted in the data set. To evaluate the performance of models, users can create a confusion matrix (Figure 1c) to compare human-labelled and predicted ratings. A confusion matrix is a performance measurement for machine learning classification that compares human-labelled and predicted ratings in a two-dimensional array. To further investigate how models make decisions, users can highlight a specific review and explore why the model prediction is different from human-labelled rating. This process is called error analysis.

In Figure 1, the highlighted review states, ‘It was fine, not great though. The meat in the shish kabob was nicely seasoned, but a bit tough’. As shown in the confusion matrix, it is labelled as negative while the model predicts it as positive. Users can create Figure 1e to explore features that drive the model decision making. In Figure 1e, each feature has a

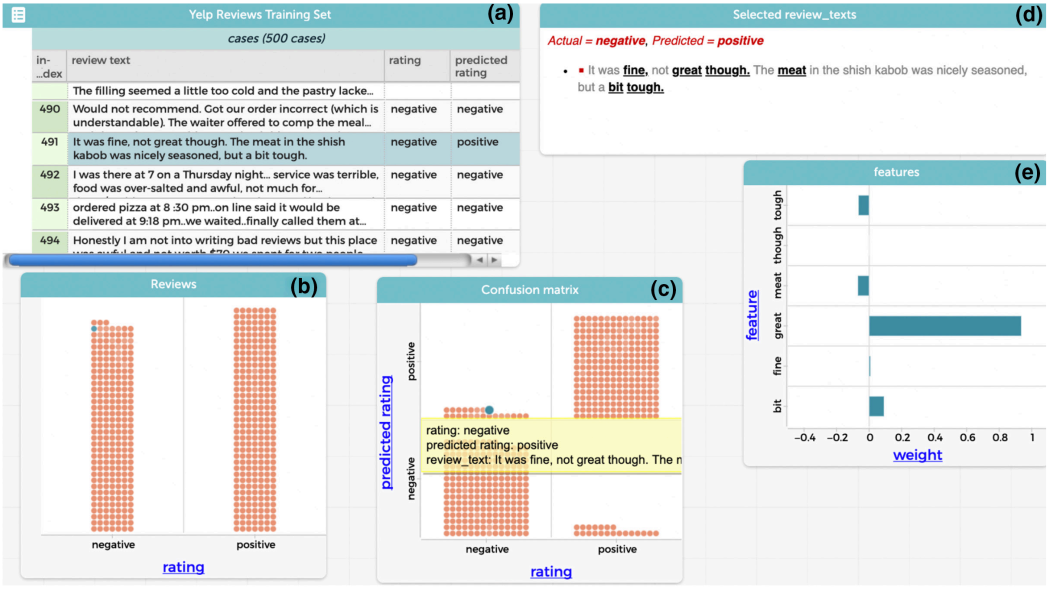

FIGURE 1 Screenshot of StoryQ: (a) yelp review data set with three columns, review texts, human-labelled rating and predicted rating. (b) Rating distribution of yelp review data set. (c) Confusion matrix. (d) Selected reviews. It shows the review that was highlighted in (c). The underlined bold texts are features considered by the model. (e) Features that the model used to make predictions of the highlighted review. (a), (b), (c), (d), and (e) are dynamically linked. When users change the highlighted review, the highlighted row in (a), highlighted dot in (b) and (c), texts in (d) and features in (e) will be updated accordingly.

feature weight. For instance, 'great' has a larger positive weight (around 0.9). If the sum of all features' weights is positive, then the predicted rating will be positive and if the sum is negative, then the predicted rating will be negative. Here, the sum is positive and thus, the predicted rating is positive. Based on the exploration, users can construct features to build a new model and retrain a new model to improve model accuracy. This is the iterative process of developing a model.

Researchers' participant-observer roles.

In this study, we took the role of participant–observer (Spradley, 1980). One of the researchers guided students to complete the learning activities while others observed the processes, collected data, took field notes and gave feedback when needed. For instance, researchers would invite students to join paired discussions when students were silent in breakout rooms. In our role of participant–observer, we had frequent interactions with students. This role allowed us to gain a comprehensive and in-depth understanding of students' learning experiences.

Data collection and analysis

We collected multiple sources of data to examine the students' experiences with the AI curriculum activities, including pre- and post-surveys, pre- and post-assessments, 28 hours of screen recordings, activity reports and semi-structured interviews. The pre-survey collected general information regarding students' background and their experiences with technology as well as their perceptions and attitudes towards AI and associated careers. The post-survey asked students' attitudes towards AI and AI-related careers. The pre- and

14678535, 2022, 5, Downloaded from https://bera-journals.onlinelibrary.wiley.com/doi/10.1111/bjet.13253, Wiley Online Library on [30/05/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

post-assessments measured students' prior knowledge about model decision making and understanding of model decision making after the classroom implementation. The recordings captured students' interactions with peers, instructors, and the StoryQ platform. While completing learning activities, students answered activity-specific questions in the activity report. At the end of classroom implementation, we conducted a semi-structured interview with one volunteer student for approximately 30 minutes to investigate data modelling experiences from his perspective.

We engaged in the following two phases of qualitative data analysis: content logging video recordings and microanalysis of data modelling using interaction analysis methods (Jordan & Henderson, 1995). We first content logged the 28 hours of screen recordings and the 30 minutes of interview recording and developed analytic memos around students' data modelling practices and processes. Then in weekly meetings, we reviewed the entire data sources to gain an overview of students' learning experiences and discussed our interpretations of each student's learning processes, focusing on social and cultural knowledge that students draw on when modelling texts. These discussions helped the research team to challenge individual assumptions and build consensus on various dimensions of model development, such as the kinds of features students considered.

Afterwards, we flagged episodes that demonstrate students' understanding of building models with texts and used interaction analysis methods to analyse these episodes. In this process, we considered discourse and tool use together to understand students' modelling texts. While gestures and gaze might provide rich information about student learning, we could not capture this piece of data as the project was implemented via Zoom and students usually turned off their camera during Zoom sessions for privacy issues and increasing Internet speed. We then sought to describe students' activities in these data modelling episodes. In this process, we iteratively analysed (Patton, 1990) analytics memos to generate the categories of activities (Hall & Stevens, 2015). We reviewed the memos and discussed common learning opportunities and challenges in these activities. We also challenged each other's perspectives on findings and interpretations throughout the analysis. In the presentation of categories of activities, we selected excerpts to show the common patterns of data modelling and to highlight voices from a range of participants. In this process, we frequently re-created models that students explored to better understand how students modelled texts and the kinds of understanding that they developed.

RESULTS

Our qualitative analysis demonstrated that students developed a detailed understanding of how AI works through modelling texts with real-world data sets. This section examines the process of developing such understandings. We illustrate four thematic categories of activities in which students engaged in modelling texts, with excerpts from students' data modelling processes.

Drawing on social experiences and cultural knowledge to create predictive features

Students helped models 'focus' on important information by creating predictive features, potentially improving the classification of positive reviews. Following a use-modify instructional approach (Lytle et al., 2019), students were guided to first use a baseline model to classify Yelp reviews as positive or negative and then modify the baseline model by adding new features with the goal of improving model accuracy. As we see it, this provides an important

14678535, 2022, 5, Downloaded from https://bera-journals.onlinelibrary.wiley.com/doi/10.1111/bjet.13253, Wiley Online Library on [30/05/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

iterative process that is in line with the socio-cultural theory guiding this work. Students are prompted to draw on their background, treating their social outlook as an important background component for creating a baseline model that will be refined as they move along.

The baseline model for both groups of students had only a single feature. For Emily and Maya, that feature was whether the review contained the word ‘great’, while Albert and Eric's baseline model was whether the review had the word ‘love’. The accuracy of Emily and Maya's baseline model was 0.66, the feature weight of ‘great’ was 1.06 and the accuracy of Albert and Eric's model was 0.53, with a feature weight of 0.39 for ‘love’. The feature weights were ranked on a scale of −1 to +1, which estimate the relative importance of the feature. An important feature would have a larger absolute value of weight than less important or irrelevant features. A larger absolute value of positive and negative weights represents more important features for classifying positive and negative reviews respectively.

After exploring this baseline model, the student groups added new features to improve the baseline model. In the process of identifying new features, both pairs examined the reviews closely and tried to identify the best features for predicting labels (ie, positive and negative). Since they have explored reviews in the data set when exploring the baseline model, their rationale for creating features might come from both understanding of the data set and their background knowledge of what might be in Yelp reviews. For instance, Maya and Emily selected whether reviews had the word ‘love’ as a new feature because they thought review writers rarely used the word ‘love’ in negative reviews (Excerpt 1).

Excerpt 1: Maya and Emily's selection of ‘love’ as a predictive feature.

> **1 Maya** (*read the question in the activity report*): What is the feature that you will be investigating?
>
> **2 Emily**: Uh, how about like…trying to think of a…I would say ‘good’, but I feel that would have around the same result as ‘great’ (*the feature in the baseline model*), (Maya, ‘nha’) just because people used it in negative reviews as well, but… I mean the word like ‘love’ would definitely be more of a positive thing because you do not really say ‘love’ if you are saying something negative.

When working in pairs, one student played the role of navigator and the other played the role of driver. The navigator read instructions and the driver shared the screen and followed the navigator's instructions to interact with the technology. The pair switched roles every session. In this excerpt, Maya was the navigator and Emily was the driver. Emily intuitively reasoned that ‘good’ was going to have positive connotations, but decided that ‘love’ seemed more definitive. Emily's rationale for ‘love’ over ‘good’ reflected a recognition of the technology's limitation: it could not interpret context. For her, ‘love’ was a word that in almost all situations represented positivity. In turn 3, Emily expressed an understanding of the concept of weight:

> **3 Emily** (*type ‘love’ for the question in the activity report, Figure 2*): Okay, we picked ‘love’ because it is not something normally you say in a negative comment, a negative rating. So it is like, more positive, probably carries more weight than the word ‘great’.

She hypothesized that the feature weight of ‘love’ would be bigger than the weight of ‘great’, which was used in the baseline model. They trained the model with this new feature and the accuracy improved to 0.68. Similarly, Albert and Eric selected the word ‘delicious’ as a new feature in their second model. As Albert reasoned, ‘it is hard to use the word

14678535, 2022, 5, Downloaded from https://bera-journals.onlinelibrary.wiley.com/doi/10.1111/bjet.13253, Wiley Online Library on [30/05/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

[delicious] in negative sentences'. His group added this feature in the model and the model accuracy (ie, 0.55) was slightly higher than the baseline model (ie, 0.53).

In these examples, we see students drawing on prior knowledge and experiences as they create predictive features for positive reviews. In particular, students gravitated towards language that carried culturally significant expressions of positivity. In the case of Albert and Eric, they reasoned that 'delicious' was not only unambiguously positive, but related directly to the experience of evaluating food. What makes 'delicious' a potentially powerful feature is its specificity: The idea of good-tasting food provides a context embedded within the word. Maya and Emily do not take the same approach, instead focusing on a straightforwardly positive and descriptive word. In each case, however, the students are drawing not only their understanding of how language works to create their models, but applying cultural patterns as they reason through how to improve on their baseline model.

Engineering predictive features to address errors

Students paid close attention to reviews in the data set to create new features for fixing misclassified reviews. As shown in Excerpt 2, Albert (driver) and Eric (navigator) used the confusion matrix (Figure 3)—a performance measurement that compares predicted and actual labels in a two-dimensional array—to explore misclassified reviews and to create new predictive features.

Excerpt 2: Albert and Eric's exploration of misclassified reviews when addressing errors.

> *(Albert clicked each dot representing reviews that were misclassified as positive by the model in the confusion matrix and read all of these misclassified reviews (Figure 3). The model was the baseline model, including one feature 'love'.)*
>
> **1 Albert**: It's usually either saying someone else loved it or they are saying that they used to love it.

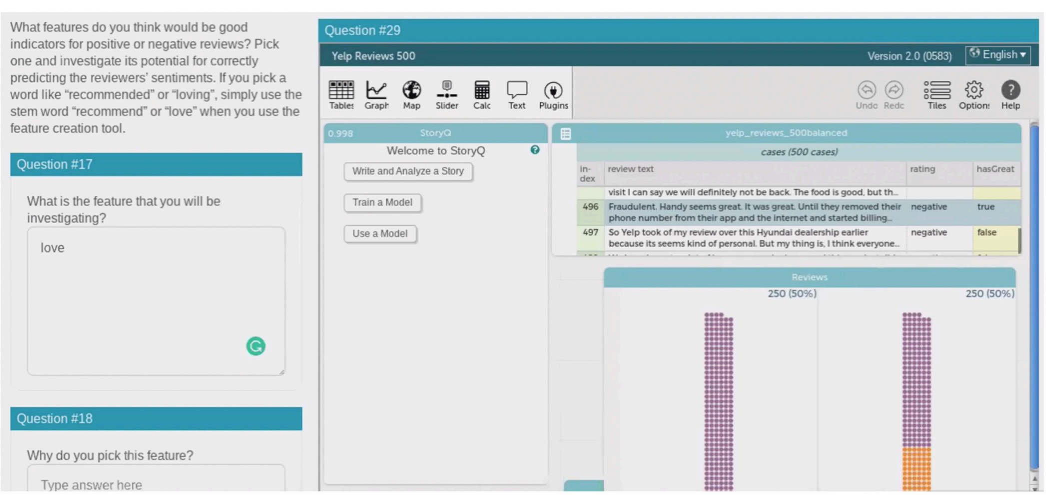

FIGURE 2 Emily (driver) shared her screen and typed 'love' in the activity report. In the dot graph, purple dots represent reviews without word 'great' and orange dots represent reviews that have word 'great'. The pair did not build a graph for word 'love' before selecting 'love' as a predictive feature.

14678535, 2022, 5, Downloaded from https://bera-journals.onlinelibrary.wiley.com/doi/10.1111/bjet.13253, Wiley Online Library on [30/05/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

2 Eric: Yeah.

(Silence for 3 seconds)

3 Researcher: Here, it asks you to add features to correct the misclassified reviews.

4 Albert: Maybe the word ‘worst’. (*three out of seven misclassified reviews included ‘worst’*).

(Albert read the misclassified reviews again.)

5 Albert: Could we use more than one feature, I mean words? *(Silence for 2 seconds)* What would be another one that we can use?

6 Eric: Maybe ‘wrong’? (*one out of seven misclassified reviews included ‘wrong’*).

7 Albert (*mouse hover the misclassified review with ‘wrong’*): Like this one?

8 Eric: Yeah.

In this example, they both proposed new features based on the misclassified reviews (turns 4 and 6). While selecting features for addressing errors, Albert selected his word, ‘worst’, based on the frequency in which it appeared in the misclassified reviews. Albert's approach reflects an in-depth understanding of feature engineering, one that begins to move beyond the kind of socio-linguist reasoning focused on selecting words that best fit with their conception of positive reviews. As Albert explained in a later interview, he found that a feature useful for the modelling task should be the one that appears frequently in the data set. Discovering those features requires not only drawing on an intuitive understanding of language, and applying this to a prior experience and knowledge of the

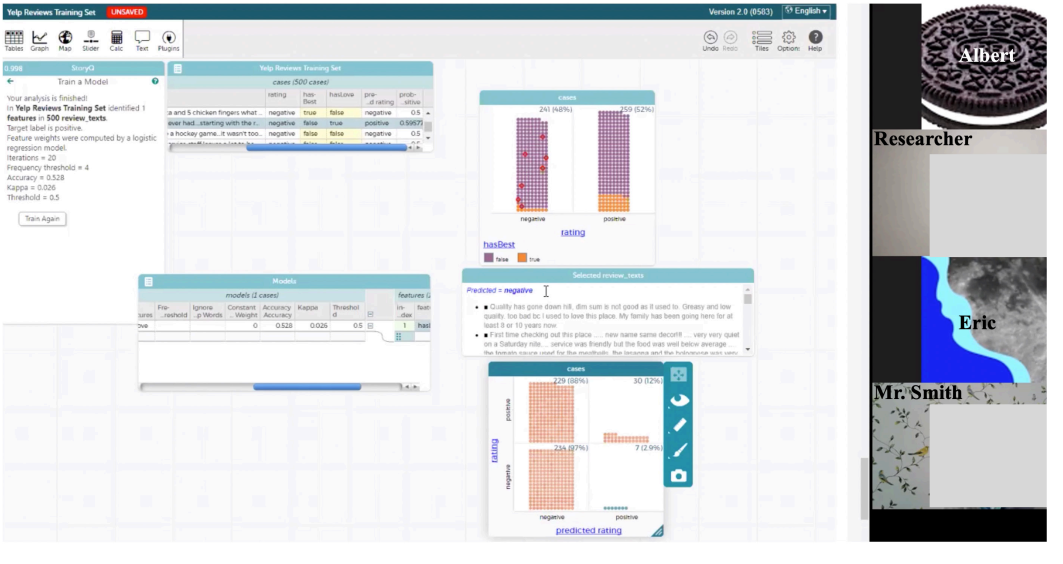

FIGURE 3 Albert (driver) shared his screen and highlighted reviews that were predicted as positive but should be negative in the confusion matrix.

14678535, 2022, 5, Downloaded from https://bera-journals.onlinelibrary.wiley.com/doi/10.1111/bjet.13253, Wiley Online Library on [30/05/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

14678535, 2022, 5, Downloaded from https://bera-journals.onlinelibrary.wiley.com/doi/10.1111/bjet.13253, Wiley Online Library on [30/05/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

way Internet reviews are written, but learning from the data set and the iterative process of modelling. This is a correct approach for improving model accuracy as addressing a frequent feature would be more likely to improve the model. For instance, if one selects a feature that only appears in one review and the model learns the feature, the model will correct one misclassified review. That is to say, a modeller should estimate the usefulness of features in feature engineering. Emily and Maya also considered features that appeared in misclassified reviews to address errors, but the pair did not express an awareness of usefulness. Students need more support to learn feature engineering practices.

Understanding that machine learning is about identifying patterns in unstructured data

When reasoning about model performance, students explained patterns learned from the Yelp review data set and described generating patterns from data as machine learning. Before taking the class, students held popular misconceptions of machine learning—in particular, the widely held view that machines learn without human intervention and in ways that are unintelligible (Boden, 2004; Chai et al., 2021; Long & Magerko, 2020). Prompted to discuss their understanding of AI before the lessons began, one student remarked that it 'gets smarter over time and figures out solutions for stuff'. Such responses indicated that students viewed machines as learning on their own over time, which is a common misconception from the public (Sulmont et al., 2019).

After exploring individual features in the unigram (or single word) model, they paid close attention to reviews in the data set and learned that models learned patterns from unstructured data. For example, in Excerpt 3, Maya (navigator) and Emily (driver) discussed 'surprising' features in a unigram model.

Excerpt 3: Maya and Emily's exploration of patterns in unstructured data.

> **1 Maya**: I think 'stop' was interesting that it had a bit more weight [than] 'clean', 'easy' and 'fast'….
>
> **2 Researcher**: Anything else?
>
> **3 Emily** (*select 'die' from the feature table, Figure 4*): 'Die' I guess.
>
> **4 Researcher**: Ha, that's strange.
>
> **5 Maya**: Because when people usually say 'stop', like, I guess, in regard to food, I guess it would be like, 'Oh, they stopped having my favorite food' or 'they stopped like doing something good'. (*As Maya spoke, Emily typed the answer in the question*).

Maya's surprise at the weight of the word 'stop' shows the importance of context or setting in which the data are generated. 'Stop' on its own may be associated with a 'stop sign' or parental command. But in the context of a food review, the word might be used as a double negative, for example, 'I cannot stop eating this food'. In turn 5, however, Maya confirmed her own hypothesis and struggled to reconcile her belief about stop's negative connotation with the weight the model attributed to the word. Yet, as the conversation progresses, Maya and Emily began to see why 'die' has a positive weight:

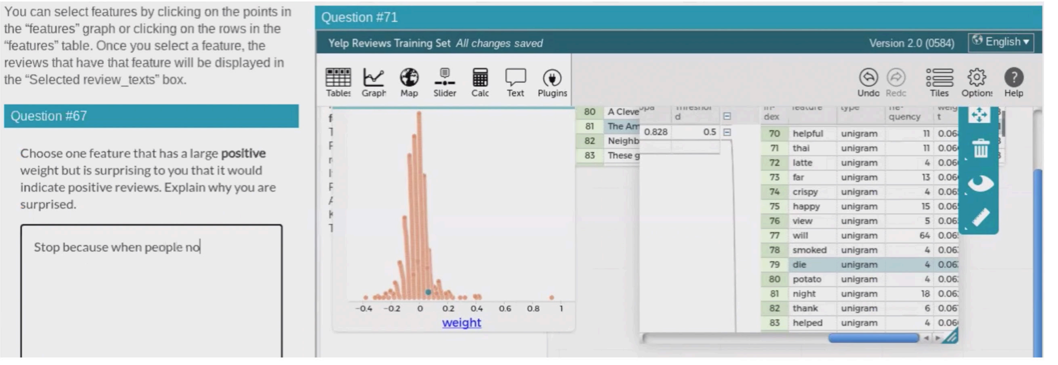

FIGURE 4 Emily (driver) shared her screen and highlighted 'die' as a surprising feature.

6 Maya: (*silence for a second*) Probably because it's not a normal word to use when you are talking about food. Like, I do not know, something like 'you could die from this', I do not know, I just feel…

7 Researcher: (*laugh*) right, it is very strange. Okay, do you want to see how people use them in the good reviews? The model learned a positive weight for these words, so they must appear a lot in the positive reviews.

8 Emily (*read the reviews*): The Americano pizza is to die for!!! (*read other reviews including the word 'die'*) They are using it in a metaphorical sense.

The positive weight of 'die', viewed in the context of the above quote, provided a powerful lesson on how patterns do not always conform to our intuitive understanding of the data. Like Maya, she began to see the power of modelling unstructured textual data as a way of moving beyond everyday language associations. They began to highlight the importance of context. In the context of food reviews, the metaphorical sense of 'die'—and the phrasing 'to die for'—is more common than its negative connotations, at least in this data set. By identifying the relevance of 'die', Maya and Emily were learning how to engage in the iterative process of identifying relevant features.

The iterative process involved in finding patterns in machine learning provides students with at least two important lessons. On the one hand, it helps to demystify how machines identify patterns and the human choices that are involved in that process. Equally important, student engagement with unstructured data reveals sociocultural dimensions of data production—after all, the reviews students analysed were written by everyday restaurant patrons. Stripping away the veneer of data objectivity, both in the way data are produced and how it is manipulated by technology, is an essential part of the process of understanding the political and cultural dynamics of technology. Through the modelling, students were gaining a nuanced and detailed understanding of the working mechanism of unstructured data and the essential role of humans in identifying patterns. We see this as an early step towards cultivating a more critical disposition towards technology.

Reasoning about noisy features in unstructured data

Students learned that model decision making was driven by multiple features and noisy features could mislead model decision making. In the process of model development, students added

14678535, 2022, 5, Downloaded from https://bera-journals.onlinelibrary.wiley.com/doi/10.1111/bjet.13253, Wiley Online Library on [30/05/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

multiple features to the model. They then compared the model performances and explored how different types and quantities of features changed the accuracies. Before adding the features to the models, they had hypothesized whether the new model would perform better and developed an understanding of model decision making by multiple features. As an example, Albert and Eric developed six models (Figure 5). Albert demonstrated his in-depth reasoning of model decision making during our interview, focusing on features that he and Eric used in the models:

> So, model 1 (*the baseline model*) just has 'love', I am pretty sure and then model 2 has just 'delicious' and model 3 has the combination of 1 and 2, (*that is to say*) model 3 has 'love' and has 'delicious', model 4 'great' and 'delicious', model 5 has 'great', 'good' and 'delicious'. Ummm, and model 6 has all of them, which is 'great', 'good', 'delicious' and 'love'. *(Researcher: what do you see about the model accuracy?)* Umm, at first, I thought you get a higher accuracy when you add more features to the model. If we continue, once we had 5 and 6, they are less than model 4 which had 2 features (*model 4 had the highest accuracy, Figure 5*). The accuracy was less. Because couple of words might lower it immediately. Umm, because model 5, we added 'good' and then model 6, we added 'good' and 'love'. And that decreased significantly.…

Albert's interview reflected his growing understanding of the way predictive models are dependent on features. As Albert explained the evolution of each model, he noted that accuracy decreased in the last two. Albert appeared to have a nascent understanding that features interacted rather than operating independently. Adding a new feature that is not correlated with an outcome variable (ie, a noisy feature) can be distracting to a model from the features that are meaningful. Yet, Albert did not fully grasp this complex concept when asked why model accuracy dropped: 'Those probably frequent usage of the word "love" or "good" and ummm, like negative sense, it probably more negative, negative usage of those two words. So when I put them into my model, they decreased the accuracy'.

For Albert, the decreased model accuracy in models 5 and 6 could be due to 'love' and 'good' being used in negative reviews. This idea could be further developed by introducing a key concept in machine learning, namely, that increased complexity resulting from more features might outweigh the added value from added features. The other three students also demonstrated an understanding of multiple features influencing model decision making. For instance, Emily explained, 'the feature needs to be useful. If it is useless, it confuses the

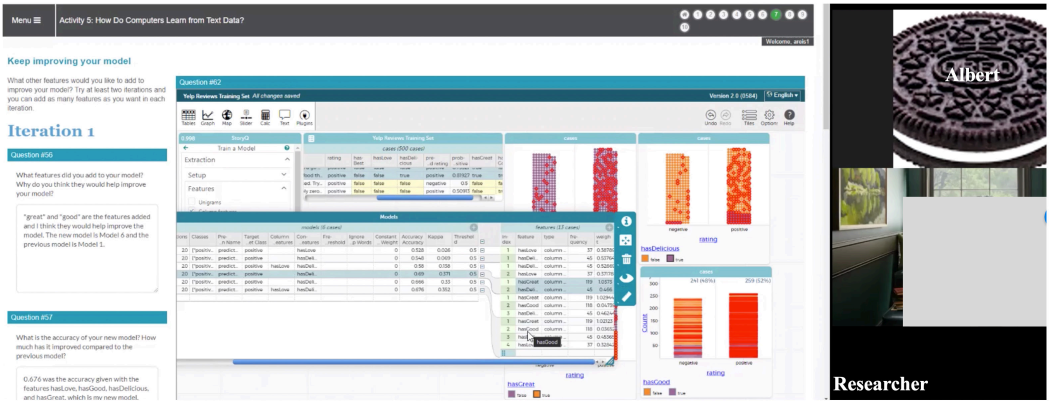

FIGURE 5 Albert shared his screen and highlighted the model with highest accuracy.

14678535, 2022, 5, Downloaded from https://bera-journals.onlinelibrary.wiley.com/doi/10.1111/bjet.13253, Wiley Online Library on [30/05/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

model'. While she did not specify the meaning of usefulness, her explanation suggests at least an inkling of understanding that model decision making, such as when having multiple features, will predict with high variance and cause less accurate results. These cases demonstrate that students were beginning to reason about noisy features in modelling unstructured data, a critical practice in building machine learning models. Perhaps more scaffolding will help students gradually move from one feature, to multiple features, as well as models with features in the thousands and millions.

DISCUSSION AND IMPLICATIONS

This study illustrates that students developed a detailed and nuanced understanding of how AI, more specifically machine learning, works. In accordance with the literature (eg, Long & Magerko, 2020; Sulmont et al., 2019), this study shows that a common preconception of machine learning from students was that machines learned on their own without human intervention. After modelling texts from a Yelp review data set, students demonstrated an understanding that modellers (students themselves) played a critical role in helping machines learn and identify patterns in data.

Drawing on the sociocultural perspective, we have also aimed to highlight the ways students drew on their cultural and linguistic background as an asset in the data modelling process. Students approached words 'good' and 'love' with both an intuitive conception about how such words are used in the context of Yelp reviews and as a way to test a hypothesis and improve the modelling performance. What became clear to students throughout these lessons is their own role in building machine learning models. Likewise, student engagement in data modelling began to build an understanding of the socially constructed nature of data. In exploring Yelp reviews, students were empowered to reflect on the way data was produced (by restaurant reviewers) and how it is manipulated through their use of technology. These findings support the current understanding of students using their cultural background as a tool to make meaning in data modelling (Shamir & Levin, 2022; Vartiainen et al., 2020). This provides a new understanding into how students can become active agents in reasoning about textual data.

Based on our microanalysis of students modelling texts we offer suggestions for future research. First, our analysis revealed that students drew on prior knowledge when creating predictive features. For instance, Emily and Maya picked the word 'love' as a feature that would help the model classify positive reviews as, from their perspective, this word was rarely used in negative reviews. The results confirmed previous findings that students were comfortable drawing their knowledge and experiences in supervised machine learning activities (Williams et al., 2019; Zhou et al., 2020). To leverage students' social and cultural background, future studies might examine other text classification tasks and strategies to create broadly inclusive learning experiences that engage participants from diverse backgrounds in expressing their cultures and personalities. For instance, given that many non-Western cultures use storytelling as an important pedagogical tool (Hooks, 2014), using narratives as a learning context can potentially engage students with the unfamiliar text mining practice and ground their reasoning in familiar domains. Students could begin with writing stories and use pre-made predictive models, referred to as classifiers, to categorize their stories in a variety of dimensions (eg, whether a character is a hero or villain). This creative writing task would allow students to bring forward their cultural and personal knowledge resources (Connelly & Clandinin, 1990) and claim strong ownership of and project their identities in the stories (Grainger et al., 2005). Additionally, in the subsequent study of preprocessing text for machine learning, students can leverage the insights of their own writings to understand structures in seemingly unstructured text

14678535, 2022, 5, Downloaded from https://bera-journals.onlinelibrary.wiley.com/doi/10.1111/bjet.13253, Wiley Online Library on [30/05/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

data. Future research can focus on exploring how students draw knowledge on their own writing in modelling texts.

Second, students in this study paid close attention to the data set in engineering features and needed more support to get familiar with feature engineering practices. This phenomenon of engaging in exploring data sets aligns with several studies in the literature (eg, Biehler & Fleischer, 2021; Sakulkueakulsuk et al., 2018). However, students encountered challenges in transforming their qualitative analysis of data sets into creating meaningful features. For example, Eric was not aware of the number of misclassified reviews that could be addressed when adding the feature 'wrong'. In feature engineering, one needs to choose features based on their impact on model performance. Feature engineering, a process of transforming unstructured data into meaningful features using knowledge about the data and the application context of the model, has the potential of helping students to understand the role and responsibility of humans in developing AI technologies as well as the importance of domain knowledge in AI application areas (Duboue, 2020; Nargesian et al., 2017). Thus, one key area for K-12 AI education research is exploring technological and instructional strategies to engage students in feature engineering practices. Third, students came to see AI as revealing patterns in data. Specifically, they learned that text classification models were trained to recognize structural patterns in the Yelp review data set. This finding shows the possibilities of integrating modelling texts into other disciplines (eg, social studies class and history) to help students to understand structured bias in AI (Lin et al., 2021) and structural patterns of discrimination. For example, students could be guided to model texts from mortgage redlining data set (ie, a real-world data set that includes textual descriptions of neighbourhoods and evaluations of whether segregated neighbourhoods were 'high risk' and should receive loans; Sadler et al., 2021). In the process of building a model classifying neighbourhoods as low-risk and high-risk, students would learn the structural patterns of racial segregation and inequality built into the data set. They would gain an in-depth view of model decision making reflecting structural patterns in the real world and be empowered to engage in discussing social justice issues in society. Thus, leveraging the affordances of text classification practices (eg, pattern identification) to help students to explore and critically evaluate social justice issues is a fruitful area for future exploration.

Lastly, when comparing models, students reasoned about multiple features driving computer decision making. This study was designed to help students understand the nuances of computer decision making. It sets a solid foundation for engaging students in reasoning about social contexts through which the data set was generated. Even within the context of this study, we could offer students opportunities to understand model decision making in rich social contexts, such as review writers in the southern and northern United States having different comments about BBQ restaurants. Furthermore, integrating rich context reasoning into modelling unstructured data (in this study, texts) could help students to critically evaluate AI technologies (Burgsteiner et al., 2016; Gresse von Wangenheim et al., 2021; Ho & Scadding, 2019). If students are presented with the context before developing a model, they are empowered to challenge what counts as useful features. Yelp reviewers, for example, reflect not only patrons' interest in the quality of a restaurant's food, but also deep concerns about customer services or access for patrons with disabilities. In general, students were aware of noisy features, but needed scaffolding to conduct in-depth reasoning about features. This finding suggests that future designs could pursue scaffolding feature exploration in social contexts that the data set is situated in.

This study holds implications for teachers to support students in modelling unstructured data and learning classification and machine learning. First, it is important for teachers to guide students to use their own cultural and linguistic backgrounds as resources for model development and more importantly reason about model decision making by drawing on such backgrounds. In addition, teachers should provide spaces for students to create their

14678535, 2022, 5, Downloaded from https://bera-journals.onlinelibrary.wiley.com/doi/10.1111/bjet.13253, Wiley Online Library on [30/05/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

own features and, at the same time, engage students in identifying patterns that models learn through these features. Furthermore, since some concepts (eg, feature and weight) and practices (eg, conducting error analysis) are challenging, students are likely to benefit from explicit instruction that helps them to understand the unfamiliar concepts and concrete steps in performing the practices. Lastly, while this study did not focus explicitly on supporting students' discussions about social contexts, we believe that these discussions are beneficial for students to gain a deeper understanding of how AI is created, how it is applied, and also its potential to perpetuate bias.

Overall, in classroom practices, teachers should stress that the process of machine learning model development involves hard choices and compromises to achieve what will always be imperfect performance. These hard choices occur at every stage, including selecting data, creating features, feeding these features into machine learning algorithms, and troubleshooting and iteratively improving the model. The resulting AI is very much what humans make it and the process of model development is creative and subjective.

In conclusion, our findings regarding student learning in modelling texts should be the starting point in research on designing instruction that promotes modelling unstructured data from real-world data sets and the understanding of how models make decisions. In particular, students should be encouraged to draw on prior knowledge and experience in modelling, creatively turn unstructured data into meaningful features and explore how AI technologies make decisions. We believe that modelling unstructured data from real-world data sets is an effective way to help students reason about computer decision making and understand important social and scientific issues about AI technologies.

LIMITATION

One limitation of the study is the small sample size. Typically, AP computer science (CS) classes in the United States are small, especially in public schools with very limited equipment, facilities, and resources. Despite the small size, the StoryQ curriculum was designed to scale. While small class size enables us to provide needed support for each student, the application of our findings in classes with a larger size and other contexts needs to be validated. The other limitation is that overall AP CS students had strong interests in CS-related topics. It would be beneficial to investigate diverse students' (eg, those without interests in CS) learning in modelling unstructured data.

ACKNOWLEDGEMENTS
This work was supported by NSF Grant DRL-1949110. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agency.

CONFLICT OF INTEREST
There are no conflicts of interest to report.

DATA AVAILABILITY STATEMENT
The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

ETHICS STATEMENT
All consent processes and forms for this study were approved by the Solutions Institutional Review Board (IRB) (https://www.solutionsirb.com/) prior to the study's implementation. In addition, the analysis was performed using non-identifiable data.

14678535, 2022, 5, Downloaded from https://bera-journals.onlinelibrary.wiley.com/doi/10.1111/bjet.13253, Wiley Online Library on [30/05/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

ORCID

Shiyan Jiang https://orcid.org/0000-0003-4781-846X

REFERENCES

AI4K12.org. (2019). *Five big ideas in AI* (Poster). https://bit.ly/ai4k12-five-big-ideas

Bapat, R., Mandya, A., Liu, X., Abraham, B., Brown, D. E., Kang, H., & Veeraraghavan, M. (2018, April). Identifying malicious botnet traffic using logistic regression. In *2018 systems and information engineering design symposium (SIEDS)* (pp. 266–271). IEEE.

Biehler, R., & Fleischer, Y. (2021). Introducing students to machine learning with decision trees using CODAP and Jupyter notebooks. *Teaching Statistics*, *43*, S133–S142. https://doi.org/10.1111/test.12279

Boden, M. A. (2004). *The creative mind: Myths and mechanisms*. Psychology Press.

Burgsteiner, H., Kandlhofer, M., & Steinbauer, G. (2016, March). Irobot: Teaching the basics of artificial intelligence in high schools. In *Proceedings of the AAAI Conference on Artificial Intelligence* (pp. 4126-4127). AAAI Press.

Chai, C. S., Lin, P. Y., Jong, M. S. Y., Dai, Y., Chiu, T. K., & Qin, J. (2021). Perceptions of and behavioral intentions towards learning artificial intelligence in primary school students. *Educational Technology & Society*, *24*(3), 89–101.

Chowdhury, G. G. (2003). Natural language processing. *Annual Review of Information Science and Technology*, *37*(1), 51–89.

Connelly, F. M., & Clandinin, D. J. (1990). Stories of experience and narrative inquiry. *Educational Researcher*, *19*(5), 2–14.

Duboue, P. (2020). *The art of feature engineering: Essentials for machine learning*. Cambridge University Press.

Enyedy, N., & Mukhopadhyay, S. (2007). They don't show nothing I didn't know: Emergent tensions between culturally relevant pedagogy and mathematics pedagogy. *The Journal of the Learning Sciences*, *16*(2), 139–174.

Estevez, J., Garate, G., & Graña, M. (2019). Gentle introduction to artificial intelligence for high-school students using scratch. *IEEE Access*, *7*, 179027–179036.

Grainger, T., Goouch, K., & Lambirth, A. (2005). *Creativity and writing: Developing voice and verve in the classroom*. Routledge.

Greeno, J. G. (1994). Gibson's affordances. *Psychological Review*, *101*, 336–342.

Greeno, J. G., & Engestrom, Y. (2014). Learning in activity. In R. K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (2nd ed., pp. 128–147). Cambridge University Press.

Gresse von Wangenheim, C., Hauck, J. C., Pacheco, F. S., & Bertonceli Bueno, M. F. (2021). Visual tools for teaching machine learning in K-12: A ten-year systematic mapping. *Education and Information Technologies*, *26*(5), 5733–5778. https://doi.org/10.1007/s10639-021-10570-8

Hall, R., & Stevens, R. (2015). Interaction analysis approaches to knowledge in use. In *Knowledge and interaction* (pp. 88–124). Routledge.

Hammerman, J. K., & Rubin, A. (2004). Strategies for managing statistical complexity with new software tools. *Statistics Education Research Journal*, *3*(2), 17–41.

Ho, J. W. & Scadding, M. (2019, June). Classroom activities for teaching artificial intelligence to primary school students. In *Proceedings of International Conference on Computational Thinking Education* (pp. 157–159). The Education University of Hong Kong.

Hooks, B. (2014). *Teaching to transgress*. Routledge.

IDSSP Curriculum Team. (2019). *Curriculum frameworks for introductory data science*. http://www.idssp.org/files/IDSSP_Data_Science_Curriculum_Frameworks_for_Schools_Edition_1.0.pdf

Jiang, S., & Kahn, J. (2020). Data wrangling practices and collaborative interactions with aggregated data. *International Journal of Computer-Supported Collaborative Learning*, *15*(3), 257–281.

Jordan, B., & Henderson, A. (1995). Interaction analysis: Foundations and practice. *Journal of the Learning Sciences*, *4*(1), 39–103.

Lee, V. R., & Wilkerson, M. (2018). Data use by middle and secondary students in the digital age: A status report and future prospects. Commissioned paper for the National Academies of sciences, engineering, and medicine, board on science education, committee on science investigations and engineering Design for grades 6–12. Washington, D.C.

Lee, Y. S., & Moon, P. J. (2020). Analysis of machine learning education tool for kids. *International Journal of Advanced Culture Technology*, *8*(4), 235–241.

Lin, C. H., Yu, C. C., Shih, P. K., & Wu, L. Y. (2021). STEM based artificial intelligence learning in general education for non-engineering undergraduate students. *Educational Technology & Society*, *24*(3), 224–237.

Lin, P., Van Brummelen, J., Lukin, G., Williams, R., & Breazeal, C. (2020). Zhorai: Designing a conversational agent for children to explore machine learning concepts. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(09), 13381–13388. https://doi.org/10.1609/aaai.v34i09.7061

Long, D., & Magerko, B. (2020, April). What is AI literacy? Competencies and design considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–16). ACM.

14678535, 2022, 5, Downloaded from https://bera-journals.onlinelibrary.wiley.com/doi/10.1111/bjet.13253, Wiley Online Library on [30/05/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Lytle, N., Cateté, V., Boulden, D., Dong, Y., Houchins, J., Milliken, A., Isvik, A., Bounajim, D., Wiebe, E., & Barnes, T. (2019, July). Use, modify, create: Comparing computational thinking lesson progressions for stem classes. In *Proceedings of the 2019 ACM Conference on Innovation and Technology in Computer Science Education* (pp. 395–401). ACM.

Marques, L. S., Gresse von Wangenheim, C., & Hauck, J. C. (2020). Teaching machine learning in school: A systematic mapping of the state of the art. *Informatics in Education*, *19*(2), 283–321.

Mobasher, B., Dettori, L., Raicu, D., Settimi, R., Sonboli, N., & Stettler, M. (2019). Data science summer academy for Chicago public school students. *ACM SIGKDD Explorations Newsletter*, *21*(1), 49–52. https://doi.org/10.1145/3331651.3331661

Nargesian, F., Samulowitz, H., Khurana, U., Khalil, E. B., & Turaga, D. S. (2017, August). Learning feature engineering for classification. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence* (pp. 2529–2535). AAAI Press.

Patton, M. Q. (1990). *Qualitative evaluation and research methods* (2nd ed.). Sage.

Petrosino, A. J. (2016). Teachers' use of data, measurement, and data modeling in quantitative reasoning. In R. A. Duschl, & A. Bismack (Eds.), *Reconceptualizing STEM education: The central role of practices* (pp. 167–180). Routledge.

Philip, T. M., & Rubel, L. (2019). Classrooms as laboratories of democracy: The role of new quantitative literacies for social transformation. In L. Tunstall, G. Karaali, & V. I. Piercey (Eds.), *Shifting contexts, stable core: Advancing quantitative literacy in higher education* (pp. 215–223). Mathematical Association of America Press.

Rosé, C. P. (2018). Learning analytics in the learning sciences. In F. Fischer, C. E. Hmelo-Silver, S. R. Golderman, & P. Reimann (Eds.), *International handbook of the learning sciences* (pp. 511–519). Routledge.

Sabuncuoglu, A. (2020, June). Designing one year curriculum to teach artificial intelligence for middle school. In *Proceedings of the 2020 ACM Conference on Innovation and Technology in Computer Science Education* (pp. 96–102). ACM.

Sadler, R. C., Bilal, U., & Furr-Holden, C. D. (2021). Linking historical discriminatory housing patterns to the contemporary food environment in Baltimore. *Spatial and Spatio-temporal Epidemiology*, *36*(100), 387.

Sakulkueakulsuk, B., Witoon, S., Ngarmkajornwiwat, P., Pataranutaporn, P., Surareungchai, W., Pataranutaporn, P., & Subsoontorn, P. (2018, December). Kids making AI: Integrating machine learning, gamification, and social context in STEM education. In *2018 IEEE international conference on teaching, assessment, and learning for engineering (TALE)* (pp. 1005–1010). IEEE.

Shamir, G., & Levin, I. (2022). Teaching machine learning in elementary school. *International Journal of Child-Computer Interaction*, *31*(100), 415. https://doi.org/10.1016/j.ijcci.2021.100415

Sorva, J., Karavirta, V., & Malmi, L. (2013). A review of generic program visualization systems for introductory programming education. *ACM Transactions on Computing Education (TOCE)*, *13*(4), 1–64.

Spradley, J. P. (1980). Doing participant observation. *JP Spradley, Participant observation*, *53*, 84.

Sulmont, E., Patitsas, E., & Cooperstock, J. R. (2019). What is hard about teaching machine learning to non-majors? Insights from classifying instructors' learning goals. *ACM Transactions on Computing Education*, *19*(4), 1–16.

Tang, D., Utsumi, Y., & Lao, N. (2019). Pic: A personal image classification webtool for high school students. In *Proceedings of the 2019 IJCAI EduAI Workshop*. IJCAI.

Tatar, C., Yoder, M. M., Coven, M., Wiedemann, K., Chao, J., Finzer, W., Jiang, S., & Rosé, C. P. (2021). Modeling unstructured data: Teachers as learners and designers of technology-enhanced artificial intelligence curriculum. In *Proceedings of the 15th International Conference of the Learning Sciences—ICLS 2021* (pp. 617–620). International Society of the Learning Sciences.

Van Brummelen, J. (2019). *Conversational agents to democratize artificial intelligence*. In 2019 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC) (pp. 239–240). IEEE. https://doi.org/10.1109/VLHCC.2019.8818805

Vartiainen, H., Tedre, M., & Valtonen, T. (2020). Learning machine learning with very young children: Who is teaching whom? *International Journal of Child–Computer Interaction*, *25*(100), 182.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

Wertsch, J. V. (1998). *Mind as action*. Oxford University Press.

Williams, R., Kaputsos, S. P., & Breazeal, C. (2021, May). Teacher perspectives on how to train your robot a middle school AI and ethics curriculum. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. *35*, No. 17, pp. 15678–15686). AAAI Press.

Williams, R., Park, H. W., & Breazeal, C. (2019, May). A is for artificial intelligence: The impact of artificial intelligence activities on young children's perceptions of robots. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–11). ACM.

Wise, A. F. (2020). Educating data scientists and data literate citizens for a new generation of data. *Journal of the Learning Sciences*, *29*(1), 165–181.

14678535, 2022, 5, Downloaded from https://bera-journals.onlinelibrary.wiley.com/doi/10.1111/bjet.13253, Wiley Online Library on [30/05/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Witten, I. H., Frank, E., Hall, M., & Pal, C. J. (2016). *Data mining: Practical machine learning tools and techniques* (4th ed.). Elsevier.

Zhou, X., Van Brummelen, J., & Lin, P. (2020). Designing AI learning experiences for K-12: Emerging works, future opportunities and a design framework. arXiv preprint arXiv:2009.10228.

Zimmerman, H. T., Reeve, S., & Bell, P. (2010). Family sense-making practices in science center conversations. *Science Education*, *94*(3), 478–505.

Zimmermann-Niefield, A., Turner, M., Murphy, B., Kane, S. K., & Shapiro, R. B. (2019, June). Youth learning machine learning through building models of athletic moves. In *Proceedings of the 18th ACM International Conference on Interaction Design and Children* (pp. 121–132). ACM.

How to cite this article: Jiang, S., Nocera, A., Tatar, C., Yoder, M. M., Chao, J., Wiedemann, K., Finzer, W., & Rosé, C. P. (2022). An empirical analysis of high school students' practices of modelling with unstructured data. *British Journal of Educational Technology*, *53*, 1114–1133. https://doi.org/10.1111/bjet.13253

14678535, 2022, 5, Downloaded from https://bera-journals.onlinelibrary.wiley.com/doi/10.1111/bjet.13253, Wiley Online Library on [30/05/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License