Education Tech Research Dev (2020) 68:2185–2214
https://doi.org/10.1007/s11423-020-09761-w

RESEARCH ARTICLE

Identifying patterns in students' scientific argumentation: content analysis through text mining using Latent Dirichlet Allocation

Wanli Xing[1] · Hee-Sun Lee[2] · Antonette Shibani[3]

Published online: 16 March 2020

Abstract
Constructing scientific arguments is an important practice for students because it helps them to make sense of data using scientific knowledge and within the conceptual and experimental boundaries of an investigation. In this study, we used a text mining method called Latent Dirichlet Allocation (LDA) to identify underlying patterns in students written scientific arguments about a complex scientific phenomenon called Albedo Effect. We further examined how identified patterns compare to existing frameworks related to explaining evidence to support claims and attributing sources of uncertainty. LDA was applied to electronically stored arguments written by 2472 students and concerning how decreases in sea ice affect global temperatures. The results indicated that each content topic identified in the explanations by the LDA— "data only," "reasoning only," "data and reasoning combined," "wrong reasoning types," and "restatement of the claim"—could be interpreted using the claim–evidence–reasoning framework. Similarly, each topic identified in the students' uncertainty attributions— "self-evaluations," "personal sources related to knowledge and experience," and "scientific sources related to reasoning and data"—could be interpreted using the taxonomy of uncertainty attribution. These results indicate that LDA can serve as a tool for content analysis that can discover semantic patterns in students' scientific argumentation in particular science domains and facilitate teachers' providing help to students.

Keywords Text mining · Latent Dirichlet Allocation · Educational data mining · Scientific argumentation

✉ Wanli Xing
wanli.xing@coe.ufl.edu

Hee-Sun Lee
hlee@concord.org

Antonette Shibani
Antonette.Shibani@uts.edu.au

[1] School of Teaching & Learning, University of Florida, Gainesville, FL 32611, USA

[2] The Concord Consortium, West Coast Office, Emeryville, CA 94608, USA

[3] Faculty of Transdisciplinary Innovation, Univeristy of Technology, Sydney, 15 Broadway, Ultimo, NSW, Australia

Introduction

Education research involves analyzing text data, such as written artifacts, essays, interview transcripts, and discourse transcripts. However, most text analyses rely on researchers' efforts, which can be prone to internal biases and inconsistencies (Yu et al. 2011). As a result, discovering patterns in large quantities of text data is logistically challenging and time consuming. Consider Massive open online courses (MOOCs). It is almost impossible to conduct thorough analyses of open-ended assignments administered in large-scale courses without using computer-aided automation. When applied properly and interpreted meaningfully, text mining has the potential to dramatically facilitate efforts to discover overarching patterns in text data produced by students.

A few studies have employed automated text analyses in educational settings, including the automated scoring of essays, constructed-response items and online forums (e.g. Beggrow et al. 2014; Chen et al., 2015; Dikli 2006; Liu et al. 2016; Shermis and Burstein 2003; Tawfik et al. 2018; Xing and Gao 2018; Xing et al. 2019a, b; Xing et al. 2019a, b; Zhu et al. 2019). Rosé et al. (2008) automated the analyses of computer-supported collaborative learning processes and developed conversational agents to support student discourse. Several studies have also used text mining to recommend resources and to facilitate discussion in online forums and live video chats (e.g. Ezen-Can et al. 2015; Abdous et al. 2012; Tane et al. 2004). Through the automatic processing of large quantities of text data, these studies aimed to provide and facilitate domain-general learning and communication.

Despite the fact that much of the student work produced in classrooms is in the form of text, few studies have investigated ways in which text mining can be used in specific domains, such as science. In this study, we used a novel, *unsupervised* text mining technology called Latent Dirichlet Allocation (LDA) to analyze the content of domain-specific texts produced by students as part of written, scientific arguments. We sought to answer two overall research questions and the sub questions in it:

- What underlying patterns would LDA identify in the support that students provided for the claims they made while completing a written scientific-argumentation assignment about the Albedo Effect? How would these patterns relate to the claim–evidence–reasoning framework? To what extent would these patterns produce correct and incorrect claims?
- What underlying patterns would LDA identify in the students' identifications of sources of uncertainty, which could weaken their arguments? How would these patterns of uncertainty attribution relate to the taxonomy provided by the uncertainty attribution framework? How would these patterns relate to uncertainty ratings?

Because the main purposes of this study were to introduce LDA and to demonstrate its potential for application in science education, we used a scientific-argumentation task as a case study. We expected that in this case study, LDA would be able to identify patterns across a large number of student-generated written arguments. The results from LDA can help teachers and instructional designers to better evaluate and support students' science learning.

Background

Scientific argumentation

Argumentation occurs in both everyday and educational settings (Kuhn 1993; Simosi 2003). Scientific argumentation is different from everyday argumentation, however, because its validity is assessed according to the norms and practices commonly accepted by the scientific community (Sampson and Clark 2008). Scientific argumentation provides both scientists and students unique opportunities to interpret data obtained during investigations in light of their understandings of the relevant phenomena (Bricker and Bell 2008) and to reflect on the limitations imposed by such investigations (Allchin 2012). The epistemic benefits of incorporating scientific argumentation into science instruction include the coordination of theory and evidence to make sense of scientific phenomena and understanding how scientific knowledge is developed and refined as new evidence and new understandings emerge (Duschl et al. 2007; Sandoval 2003). Scientific argumentation has also been promoted as one of eight scientific practices that should be implemented in science classes (National Research Council 2012; NGSS Lead States 2013).

Scientific argumentation is carried out by means of language (Walton et al. 2008), rhetorically (Sampson and Clark 2008), or dialogically (Clark and Sampson 2008) using commonly recognized elements (Toulmin 1958), such as:

- A *claim* that answers the question driving an investigation.
- *Data* that support the claim.
- *Warrants* based on knowledge available to the investigator that explain how the data support the claim.
- *Backing*, or the select collection of established scientific facts from which the warrants are drawn.
- *Qualifiers*, which indicate the strength of the claim given the evidence and the backing.
- *Conditions of rebuttal*, which specify circumstances in which the claim may be inapplicable because of methodological, conceptual, or contextual limitations.

Most of the analytic frameworks that can be used to analyze written scientific arguments focus on the claim–evidence–reasoning expressed in the claim, the data, the warrants, and the backing (Clark and Sampson 2008). These frameworks assess how well students coordinate theory and evidence. No studies have used qualifiers and conditions of rebuttal to analyze written arguments, however, even though investigations should critically evaluate the evidence—data never support scientific claims with absolute certainty. Instead, existing research has conceived of them as counterarguments made by multiple students or groups of students focusing on flaws in claim–evidence coordination (Erduran et al. 2004).

Lee et al. (2014) recently proposed an uncertainty-infused framework of scientific argumentation that can be used to assess students' written arguments using all six of Toulmin's (1958) elements. On this framework, qualifiers are understood as expressing the degree of uncertainty in a claim, while conditions of rebuttal are understood as attributing sources of uncertainty rooted in epistemic or ontological limitations. The Rasch modeling of students' written arguments showed that the framework can be used to interpret *uncertainty* and to link claims and evidence with reasoning.

Uncertainty has solid theoretical foundations in self-regulated learning and particularly related to monitoring accuracy and hard-easy effect. Monitoring accuracy (over/

underconfidence) is able to influence greatly on students' learning and memory (Dunlosky and Rawson 2012). Studies have shown that learners' confidence is able to predict their test results and their realistic with their goals (Huff and Nietfeld 2009). In the meantime, according to the hard-easy effect, individuals tend to be overconfident on hard tasks and underconfident on easy tasks (Stone 2000). Under or over-confidence can be explained from their uncertainty level. Therefore, examining the students' uncertainty level in their argumentation can give teachers' insights to help students' regulate their learning process.

So far, the assessment of students' written scientific arguments has relied on human judgments. Most human-coding approaches have used Toulmin's (1958) characterization of argumentation to determine which aspects of scientific arguments to assess. These studies have generally used qualitative approaches to determine the nature of argumentative discourses (Aufschnaiter et al. 2008; de Vries et al. 2002) or to identify patterns inherent in written arguments (Berland and Reiser 2009; Sandoval and Millwood 2005). For example, Erduran et al. (2004) used a cumulative coding scheme on which scores increased as additional structural elements were added. Sadler and Fowler (2006) developed a rubric consisting of claims "without justification," "with no valid grounds," "with simple grounds," "with elaborated grounds," and "with elaborated grounds with a counter-position." However, the time- and effort-intensive nature of qualitative approaches requires that analyses be conducted on relatively small samples, and this can limit the generalizability of the results.

Text mining

Text mining focuses on automatically identifying and extracting interesting and non-trivial information from unstructured text (Feldman 1995), and it uses methods from information retrieval, machine learning, data mining, statistics, and computational linguistics. Unlike the traditional mining of structured databases or XML files, text mining can handle unstructured or semi-structured data, including e-mails, full texts, and HTML files. Text mining begins with the preprocessing of textual data and ends with the storing of extracted information in a data structure suitable for retrieval. Most text mining methods assume that a text document can be represented as a set of words (i.e. a *bag-of-words*) (Hotho et al. 2005). A vector representation for a text document is constructed for an identified set of words. The importance of each word in the document is determined by assigning it a numerical importance value.

Education research has used qualitative paradigms (such as grounded theory and content analysis) to analyze texts without the explicit aid of computer algorithms. Therefore, questions related to text mining's unique contribution to and compatibility with qualitative research methodologies have been raised and actively discussed in the literature (Janasik et al. 2009; Yu et al. 2011). Yu et al. (2011) described three elements shared by text mining and qualitative research. First, text mining, like grounded theory, aims to iteratively refine the theoretical or analytical framework(s) that the researcher is applying to the text data by adding, deleting, and revising initial patterns, observations, and categories. Second, text mining, like content analysis, identifies common themes by processing natural language. Third, the quality and validity of the results of text mining are subject to the same criteria that are used to evaluate qualitative research, including reliability and consistency. There is no doubt that human inspection and insights are needed to interpret the categories and patterns generated by automated text mining algorithms.

Some educational applications of text mining can be found in the literature. Akçapınar (2015) used text mining to identify similarities in text documents to reduce plagiarism in online writing assignments. Lin et al. (2009) used text mining to distinguish different genres of threads in online discussions. To measure civic scientific literacy in the media, Tseng et al. (2010) used text mining to draw concept maps for news stories. Hung (2012) processed 689 refereed publications using hierarchical agglomerative clustering to identify longitudinal trends in e-learning research. Abdous and He (2011) analyzed live video streams that featured textual data using various clustering and classification algorithms to identify students' technology-related problems. Chen et al. (2008) produced an e-learning-domain concept map by mining academic articles. Tane et al. (2004) used *k*-means clustering to group e-learning resources and documents according to similarities in their contents. Some studies have examined the automated scoring of essays (e.g. Shermis and Burstein 2003; Dikli 2006). LDA also has applications in education. Ezen-Can et al. (2015) applied LDA to a MOOC discussion forum to group similar posts into clusters and to investigate the structures of these clustered posts. Southavilay et al. (2013) used LDA to examine the evolution of topics in collaborative writing processes. Similarly, Chen (2014) used LDA to track topics and changes in a collaborative discussion context.

One subfield of text mining focuses on analyzing argumentative texts—including essays, legal documents, and research publications—in the sciences. In this subfield, text mining applications are used to identify in different argumentative texts various structural elements of argumentation, including claims, warrants/reasoning/backing, rebuttals, and data/evidence (Zhang and Litman 2015). Instead of being used to identify domain-specific conceptual tendencies or difficulties, the findings of such analyses can be used to compare different text documents or to track revisions. The argumentative texts analyzed in structural text mining are relatively long; unsupervised text mining is not commonly used to analyze the contents of short argumentative texts, i.e. texts approximately three to seven sentences in length. No studies were found that explored whether text mining (including LDA) can be used to find semiotic patterns in students' scientific arguments.

Research context

Framing scientific argumentation

Data on students' written scientific argumentation were collected as part of a NSF-funded project called "High-Adventure Science (HAS)." These data are available at https://concord.org/high-adventure-science/. The HAS project created six interactive curriculum modules related to Earth and space science for high school and middle school students. The project contextualized students' investigations into current scientists' inquiries, such as "What is the future of Earth's climate?" and "What are our choices for supplying energy for the future?".

In the HAS modules, scientific argumentation was systematically incorporated into science activities. Prior to completing writing tasks involving scientific argumentation, students analyzed and interpreted scientific data collected by scientists or generated by manipulating models. Students also read scientific materials to gain the background necessary to interpret the data. Students then responded to four-part scientific-argumentation prompts designed to help them develop their arguments. They were asked to:

- Make scientific claims by selecting an answer from multiple choices (*claim*).
- In open-ended responses, explain claims using evidence and theory (*explanation*).
- Express their levels of uncertainty regarding their explanations for their claims using a five-point Likert scale ranging from "not at all certain" (1) to "very certain" (5) (*uncertainty rating*).
- In open-ended responses, identify sources of uncertainty (*uncertainty attribution*).

As Fig. 1 shows, the students' responses to the "claim" and "uncertainty-rating" prompts were numerical, while their responses to the "explanation" and "uncertainty-attribution" prompts were textual. Text mining was applied to the open-ended explanations and uncertainty attributions that the students produced.

These four-part scientific-argumentation prompts were developed to elicit uncertainty-infused scientific argumentation (Lee et al. 2014). The framework of uncertainty-infused scientific argumentation emphasizes the fact that scientific claims based on data cannot be

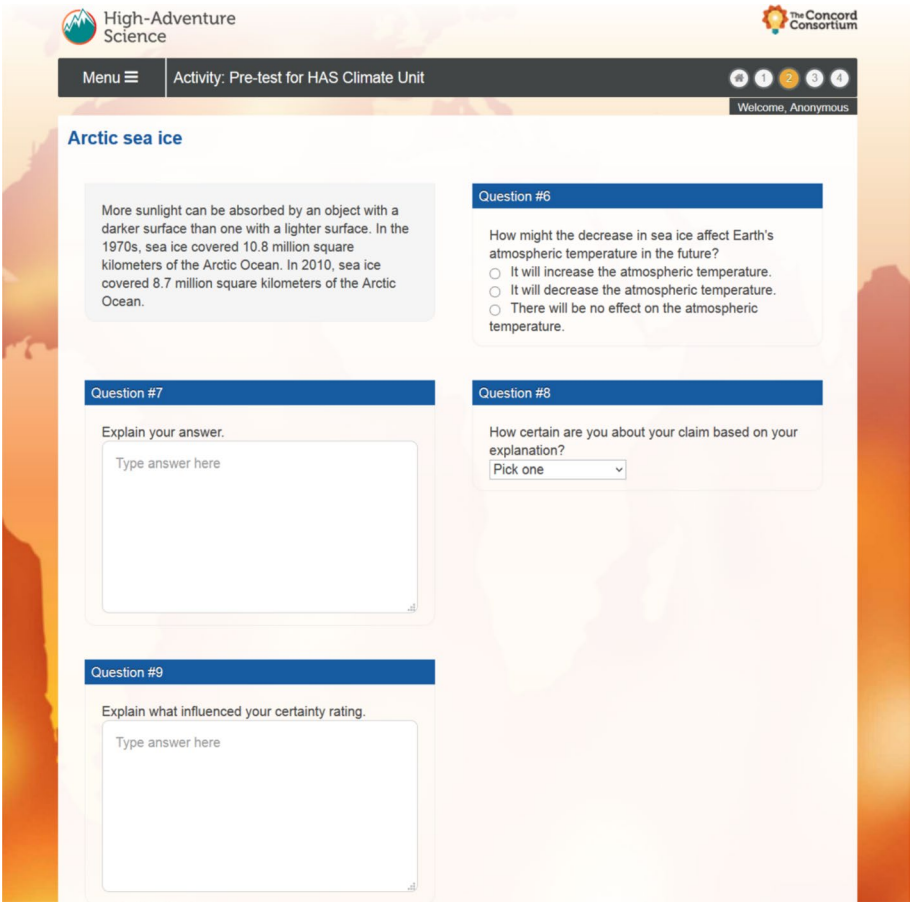


Fig. 1 The four-part scientific-argumentation task related to the Albedo Effect in Question #6 is the claim, Question #7 is the explanation, Question #8 is the uncertainty rating, and Question #9 is the uncertainty attribution

made with absolute certainty (Staley 2014) because of the ways in which scientific investigations produce data and the ways in which scientific knowledge is used to interpret data. The "claim," "explanation," "uncertainty rating," and "uncertainty attribution" prompts were intentionally separated because students have difficulty distinguishing among claims, data/evidence, and warrant/reasoning when engaging in unguided free writing (Berland and Reiser 2009) and they may not include uncertainty related to their claims when they are not explicitly asked to do so.

Scientific-argumentation task on the albedo effect

This task addressed how changes in albedo (i.e. the reflection of light off of the Earth's surface) triggered by losses of sea ice can affect Earth's temperature. This argumentation task was delivered online via the HAS module "What is the future of Earth's climate?" ("Climate change" for short). The students' responses were collected electronically. As Fig. 1 shows, the task provided data—e.g. "In the 1970s, sea ice covered 10.8 Million square kilometers of the Arctic Ocean. In 2010, sea ice covered 8.7 Million square kilometers of the Arctic Ocean." In addition, the task included the information necessary for the students to reason about the color of the Earth's surface and the absorption of the sun's radiation by the Earth's surface. Students were thus provided both the data and the knowledge required to form their arguments. By analyzing the students' written responses to the "explanation" and "uncertainty attribution" prompts, we discovered how the students reasoned in their explanations using particular pieces of data and how they thought about sources of uncertainty in their uncertainty attributions.

Over a 3-year period, the climate-change module in which the albedo-effect argumentation task was embedded was employed in 11 U.S. states by 24 middle and high school teachers and 2472 of their students. These teachers were recruited via Earth-science-listserv emails and conferences for science teachers. Of the students, 47.4% were male, 8.0% were English Language Learners, and 52% regularly used computers for school work. The average age of the students was 14.06 (SD = 1.80). In this study, LDA was applied to the students' open-ended textual responses to the "explanation" and "uncertainty attribution" prompts. Claims were coded as "correct" or "incorrect." Uncertainty ratings were coded from 1 (not at all certain) to 5 (very certain).

The analytic framework used in this uncertainty-infused scientific-argumentation task separated explanations of claims from identifications of sources of uncertainty. We used claim–evidence–reasoning to group the explanations into the following categories:

- Claim without explanation
- Partial explanation without details regarding data or reasoning
- Explanation containing elaborated data but not reasoning
- Explanation containing elaborated reasoning but not data
- Explanation containing both elaborated data and elaborated reasoning

In addition, we developed a taxonomy of uncertainty attributions that included the following:

- *Introspective confidence*, which indicated the student's personal confidence or certainty level.

- *Internal rationale*, which was based on the student's personal knowledge and experience.
- *External source acknowledgement*, which generically referred to data or knowledge
- *External scientific disposition*, which articulated the scientific data or knowledge that the student used in support of their claims.
- *External scientific limitation*, which described the theoretical, methodological, measurement, and interpretive limitations inherent in the data collection and analysis.

We used these above two analytic frameworks when interpreting the topics that LDA identified. These topics are described in the results section.

Methods

Overview

To go beyond word frequency count, a text mining method, Latent Dirichlet Allocation (LDA), was proposed to dig into the content in order to automatically identify the general topic patterns in students' explanation and uncertainty argumentation. To facilitate the understanding of the discovered topics, we create a dynamic visualization of LDA developed by Sievert and Shirley (2014). These visual analytics help us discover the meaning of each topic, examine the prevalence of each topic, and estimate how these topics relate to each other. We then characterize each of the identified topic by examining the most representative topical words and present the typical argumentation examples. In order to show the different identified topics by LDA influence students' actual scientific argumentation performance, statistical analysis was performed to examine student performance difference in students' explanations (Chi-Square independence test) and uncertainties (one-way ANOVA). Students' argumentation performance ratings on explanations and uncertainties were rated manually by experts in science learning.

Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a probabilistic model commonly used for topic modeling in natural language processing. It is essentially an unsupervised clustering algorithm that automatically identifies topics common among text documents. Clustering algorithms like LDA assume that: (1) a text document is comprised of several topics, (2) each topic consists of several words, and (3) the probability of each topic appearing in a given text document can be calculated. Topics are identified in large sets of text documents (i.e. data corpuses) by computationally examining important words that appear in both a given text document and throughout the entire corpus.

The study was conducted in five steps. The first three steps involved identifying patterns of topics by applying LDA to the argumentative responses that the students provided in their explanations and uncertainty attributions. **Step 1** involved pre-processing to remove noise from the textual data. A series of standard pre-processing techniques were performed, including removing all non-letter symbols, numbers, punctuation marks, stop words, and stemming. This preprocessing resulted in a data corpus in the form of a "bag-of-words" that took into account the number of times that words occurred but not the order in which

they occurred. In this step, we also used our corpus to show the general frequency with which students used particular words in their explanations and uncertainty attributions.

Since there was no definitive prior knowledge on how many topics should be identified in students' explanations and uncertainty attributions, **Step 2** was necessary to computationally determine the optimum number of topics (K). The aim was to maximize the differences among the computationally discovered topics and to minimize the differences within each topic. We combined the optimally-computed K with the analytic frameworks we adopted for this study. Specifically, claim-evidence-reasoning framework was adopted to inform the decision on the number of topics on explanations, and uncertainty-attribution taxonomy was used to inform the topic number decision on uncertainty arguments. Particularly, a scientific argument expert with more than 15 years of experience will use the framework to make an informed decision.

In **Step 3**, we used LDA to automatically cluster the scientific arguments into different topics. At the same time, the words that would best represent each topic were identified. These words were used to determine what the topic would mean in the context of the argumentation task. Since step 3 identified the topics and patterns, it yielded the most important results. Appendix details the process by which LDA was implemented. In total, 2472 "explanation" responses and 2472 "uncertainty-attribution" responses were separately analyzed. To further examine the discovered topics, we created a dynamic visualization of LDA based on Sievert and Shirley (2014). This visualization helped us to discover the meaning of each topic, to determine the prevalence of each topic, and to estimate how the topics related to each other. We then described each topic by identifying the most representative topical words and compared them with typical examples of argumentation containing them. We interpreted the identified topics using two analytical frameworks. We used claim–evidence–reasoning for explanations, and we used the uncertainty taxonomy for uncertainty attributions.

In **Step 4**, to validate the results of topic modeling, we conducted qualitative analysis to further show the validity of the derived topics. As there is no gold standard list of topics can be generated to compare to newly discovered topics, many studies apply a variety of quantitative measures of model fit such as perplexity or held-out likelihood to evaluate the topic models. These metrics are useful for evaluation of the predictive model; however, do not address the more explanatory goals of topic modeling. That is, how these generated topical words represent the latent space and to what extent users can understand and make sense of the topical words.

In this study, a qualitative evaluation was conducted on the topic models called topic intrusion (Chang et al. 2009). Topic intrusion examines whether the topics produced via topic modeling matches human judgments of the topics addressed in the student responses. This provides an evaluation of the latent space showing whether the topics produced describe the student responses well. A number of student responses are randomly selected from the whole dataset. A document *p*, or in our context, a student response, may belong to one or several topics. The task is to identify how the topics t_i discovered match this student response. Two doctoral students with extensive educational research background conducted this analysis together to improve its reliability.

In **Step 5**, we conducted additional statistical analyses. To examine how the LDA-identified explanation topics related to correct and incorrect claims, we conducted a Chi-Square independence test on percentages of topics making correct claims vs. incorrect claims across topics identified by LDA. To examine how the uncertainty ratings were linked to the LDA-identified uncertainty-attribution topics, we conducted a one-way ANOVA on the uncertainty ratings in the uncertainty attributions for all of the topics identified by LDA.

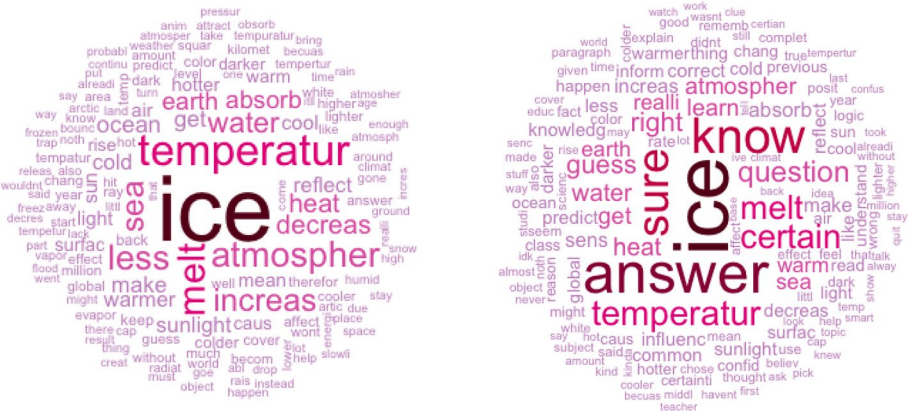
In sum, the results of Step 1 function as a descriptive visual statistic for general word usage. The results of Step 2 (the K-values) were needed to justify the number of topics that would be extracted from the text corpus. The results of Step 3 and Step 5 addressed research questions 1 and 2, which related to topic identification and the interpretations extracted from the students' explanations and uncertainty attributions.

Results

Step 1: General word usage

Figure 2 presents word-cloud maps generated from the explanations and uncertainty attributions after Step 1 was completed. These overview the general word usage in the students' explanations and uncertainty arguments. Words in larger font sizes occurred more frequently than did those in smaller font sizes. The words in these cloud maps became a list of words to transform each student's responses into a vector.

As Fig. 2 shows, "ice," "melt," and "temperature" are three of the words that occur most frequently in both word-cloud maps. However, the other frequently occurring words differ between "explanation" and "uncertainty attribution" responses. For instance, the students' explanations included words associated with the domain of climate science, like "atmosphere," "water," "heat," "increase," "decrease," and "sea." Students' uncertainty attributions included words necessary to express uncertainty, such as "answer," "know," "question," "certain," and "sure." These maps therefore indicate that the students treated the "explanation" and "uncertainty attribution" prompts differently, focusing on scientific reasoning in responding to the "explanation" prompt and on uncertainty elaboration in responding to the "uncertainty attribution" prompt.



(a) Explanations (b) Uncertainty Attributions

Fig. 2 Word-cloud maps

Step 2: Optimal topic numbers (K values)

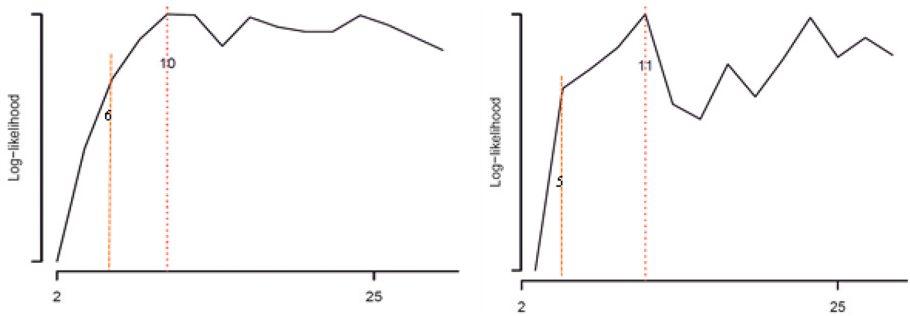
To determine how many topics should be identified in the students' explanations and uncertainty attributions, we used the log-likelihood method, on which a range of K values from 2 to 30 were compared (see Fig. 3). The larger the log-likelihood value, the better the topic grouping to extract information from the corpus. As Fig. 3 shows, the highest log-likelihood was associated with K = 10 for the explanations and K = 11 for the uncertainty attributions. Figure 3 also shows that the log-likelihood values increased sharply between K = 2 and K = 6 for explanations and from K = 2 to K = 5 for uncertainty attributions. After these drastic changes, however, the changes in the log-likelihood values became less pronounced. As a result, we selected six topic groups for the explanations ($K = 6$) and five topic groups for the uncertainty attributions ($K = 5$) along with the insights from scientific argument experts.

Step 3: Research question 1 on explanations

Visualizing the LDA results

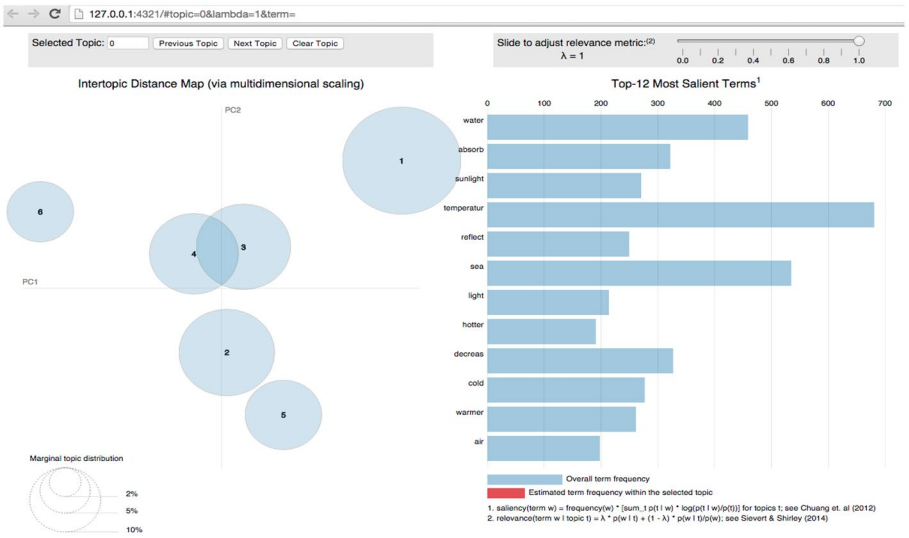
Figure 4a presents dynamic visual analytics for the explanations that the students wrote to justify their claims about the impact of sea-ice loss on the global climate. The left side of Fig. 4a shows a topic map with six circles. Since each circle corresponds to a topic identified by the LDA algorithm, these six circles represent the six topics most touched upon by the students in their explanations. Note that the number of circles (and thus the number of topics) resulted from the analysis of the log-likelihood graph in Fig. 3a, after which K = 6 was chosen. The locations of the centers of the circles represent the distances between the topics as determined by the LDA algorithm. The closer are two circles, the more closely are related the topics that the circles represent. The LDA algorithm automatically assigned a number to each topic—Topic 1, Topic 2, etc.—according to the prevalence of the topic within the entire corpus. The largest circle represents Topic 1, which was the most prevalent among the explanation-data corpus. The smallest circle represents Topic 6, which was the least prevalent among the data corpus.

The bar chart on the right side of Fig. 4a lists the top 12 words that were most relevant to the six topics discovered by the algorithm. "Water," "absorb," "sunlight," "temperature,"

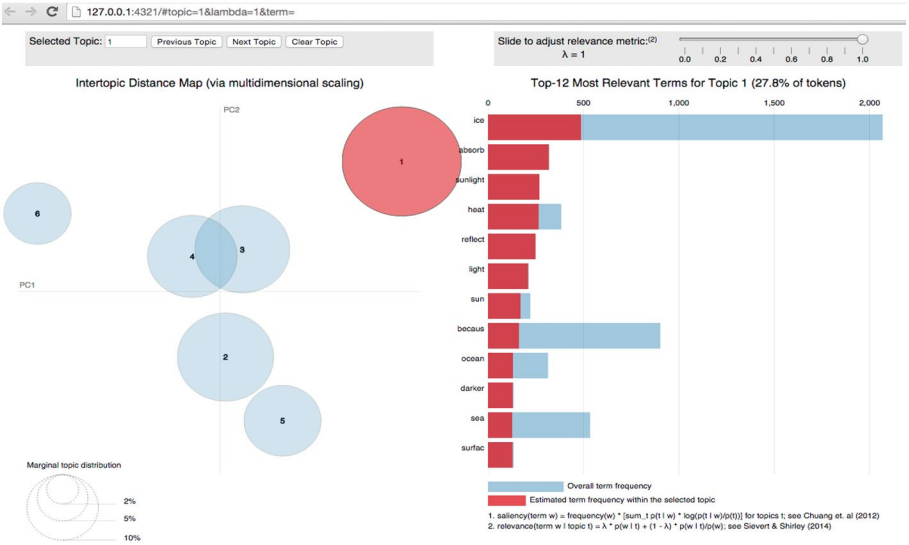


(a) Explanations **(b)** Uncertainty Attributions

Fig. 3 The log-likelihood graphs used to determine the K values



(a) The six topics identified in the explanation responses



(b) Topic 1 is highlighted

Fig. 4 The LDA results for students' explanations are visualized. **a** The topic map on the left shows the relative positions and sizes of the six topic groups (K = 6), and the bar graph on the right shows the top 12 most relevant words for the six topic groups. **b** Topic 1 highlighted in red with a list of most relevant words belonging to Topic 1 on the right

"reflect," "sea," and "light" were the most important words in distinguishing the six topics. Even though Fig. 2a indicates that the word "ice" was the most prevalent in the students' explanations, "ice" was not listed as one of the top 12 most important words for

topic identification. "Ice" was not very useful in distinguishing among topics because "ice" appeared in almost all of the students' explanations. The length of the bar for each word represents the number of student explanations that included that word. The longer the bar, the more frequently the word appeared in the students' explanations. "Temperature" occurred 700 times in the 2472 explanations, and "water" appeared 460 times.

When a particular circle (topic) is selected from the topic map on the left, the right side lists the top 12 most prevalent words in that topic and describes how these words were distributed within the data corpus containing that topic (the red bars) and within the entire data corpus (the blue bars). In Fig. 4b, Topic 1 was chosen. By comparing for each word the size of the red bar to that of the blue bar, it can be determined whether the word was present only in a given topic. When such a comparison is made for Topic 1, it is clear that only Topic 1 included "absorb," "sunlight," "reflect," "light," "darker," and "surface." We can thus tell that students were writing about sunlight absorption, light reflection, and a darker surface, which are key to explaining the albedo effect.

Characterizing the identified topics using the most-representative topical words

In Fig. 4a, Topic 1 is represented by the largest circle and is positioned the farthest away from the other five topic circles. Topics 2, 3, 4, and 5 are located close to one another, while Topic 6 is isolated from the rest. As this visualization indicates, topics 1 and 6 had characteristics different from the other topics, while topics 2, 3, 4, and 5 shared some salient words among them.

Table 1 lists the number of explanations that included a given topic. Note that an explanation can include more than one topic. In Table 1, the total number of instances may not be exactly the same as the original number of instances (2472) due to preprocessing. The examples in Table 1 are direct quotes from students and may contain typographical or other errors. As Table 1 shows, 817 explanations contained words salient in Topic 1, 721 explanations contained words salient in Topic 2, and so on. Table 1 also includes the top 12 most important words used in a given topic. Most words exclusive to the topic are marked with an asterisk. For example, the most important words in Topic 1 included "ice," "absorb," "sunlight," "sun," "light," "heat," "reflect," "ocean," "darker," "sea," and "surface." Of these words, "absorb," "sunlight," "reflect," "light," "darker," and "surface" occurred exclusively in Topic 1. The words identified as exclusive to a given topic were used to interpret what the topic meant in terms of claim, evidence (data), and reasoning. For the albedo-effect scientific-argumentation task, a scientifically valid argument included:

- A claim: the global temperature will increase.
- Data: the amount of sea ice decreases (over time) or, more specifically, it shrank from 10.8 Million square kilometers in the 1970s to 8.7 Million square kilometers in 2010.
- Reasoning: because ice reflects sunlight, decreases in sea ice darken the Earth's surface and cause it to absorb more sunlight.

From this, we can easily interpret Topic 1 as expressing the reasoning required by the argumentation task. An example response that uses many of the top 12 most important words is given below:

> Temperature will increase because since *light*er *surfaces absorb* little *sunlight* and since there is less *sea ice* to *absorb* the *sunlight* the temperature will increase

Table 1 Explanations: words are presented in order of importance for each of the six topics

	Number of occurrences (N = 2472)	Most important words	Examples	Characterization
Topic 1	817	Ice, absorb*, sunlight*, heat, reflect*, light*, sun, because, ocean, darker*, sea, surface*	• The ocean is dark so it could attract more heat • With the melting of the sea ice, there will be a greater chance for the surface of the earth to absorb more light, and therefor heat • Lighter surfaces reflect more sunlight, so more sunlight would be reflected back into the atmosphere	Full reasoning
Topic 2	721	Ice, temperature, atmosphere, because, melt, increase, sea, decrease, earth, cool, cold,	• The ice melt will cause the atmosphere to be warmer • The decrease in sea ice will increase earth's atmospheric temperature because with the ice melting, it's not going to be as cold • If the ice is melting, then it must therefore be a higher temperature in the climate than it was before	Data to claim
Topic 3	707	Water, ice, melt, because, sea, temperature, ocean, make, air, heat, cold, cause	• There will be more rain because it is evaporating the water causing temperatures to decrease • The more water there is, the more water will evaporate into the air. If there is more water in the atmosphere, it would probably make it cooler • Ice melting has will effect on the atmospheric temperature. When the ice melts the water in air humidity will go up. When the Earth is humid it creates a warmer temperature	Data to claim plus intermediate reasoning

Table 1 (continued)

	Number of occurrences (N = 2472)	Most important words	Examples	Characterization
Topic 4	640	Ice, because, melt, warmer, hotter, make, cold, air, cool, colder, earth, temperature	• The ice is melting, causing the ocean to warm which melts more ice • Ice helps keep the temperature cool. Less ice will make it be warmer • Because there is so much ice that it cools everything down in our atmosphere the more ice that is gone the hotter it will get	Data to claim plus wrong reasoning
Topic 5	576	Ice, temperature, sea, decrease, because, cover*, ocean, million*, atmosphere, increase, square*, year*	• Well, in the 1970s, there was sea ice covered about 10.8 Million square kilometers as in the year of 2010, the sea ice covered 8.7 Million square kilometers of the Arctic Ocean • The ice cover in the Arctic Ocean went down in kilometers from the 1970′s to 2010	Data
Topic 6	482	Because, ice, global*, guess*, warm, change*, answer, don't*, water, melt, increase, cool	• I'm just guessing • I don't know • I'm am not positive about my prediction • The temp will increase • Why this is happening because of global warming	No data, No reasoning

*Most words exclusive to the topic are marked with an asterisk

since there are more *darker surfaces* exposed to *absorb* the *sunlight*, increasing the temperature.

Table 1 lists three other student explanations that included Topic 1. We can see consistency in these students' explanations because they each included the most important words that LDA identified for Topic 1.

The proximity of topics 2, 3, 4, and 5 signals similarities between them: they all have to do with decreases in or the melting of ice. Note that "ice" and "*melt*ing" are two of the most important words in all four of these topics. However, subtle differences can be noted. As Table 1 shows, Topic 5 differs from topics 2, 3, and 4 because "cover," "million," "square," and "year" are exclusive to it. This indicates that Topic 5 is related to the direct citation of the data that appeared in the argumentation task itself: "In the 1970s, there was sea ice covered about 10.8 Million square kilometers as in the year of 2010, the sea ice covered 8.7 Million square kilometers of the Arctic Ocean." Some of the explanations in Topic 5 directly cited the data in linking decreases in sea ice to temperature decreases (which is an incorrect claim)—e.g. "I think that since the ice is decreasing the temperature will drop, too." Topic 2 is closer to Topic 5 than are Topic 3 or Topic 4 because topics 2 and 5 share "sea," "ice," and "decreasing." However, most Topic 2 responses connected decreases in sea ice (the data) directly to temperature increases (supporting the correct claim). Note that both Topic 2 and Topic 5 responses do not involve scientific reasoning. This indicates that being able to cite data does not always result in choosing a correct claim or providing scientifically elaborated reasoning.

Topics 3 and 4 differ from topics 2 and 5 because they provide additional information beyond that regarding decreases in sea ice and changes in temperature. Topic 3 responses used water as a means to formulate reasoning that could be valid or invalid, depending upon what role water plays in changing global temperature. For instance, "Melting ice creates more water, which evaporates to make the air cooler because evaporated water is cold" (this is an example of invalid reasoning). Some students' explanations indicated that water evaporation warmed the air because high humidity is directly associated with warm air (this is an example of invalid scientific reasoning). Other explanations indicated that water vapor is a greenhouse gas (this is an example of valid scientific reasoning). Topic 4 responses used ice being cold as a main mechanism to create warmer air because ice melts as the ocean or the air warms (this is an example of correlational, valid reasoning for feedback mechanism) or to create colder air because melting ice makes the ocean cold (this is an example of invalid reasoning). Topic 3 and Topic 4 responses overlap because they both use the coldness of ice to explain the cooling of air. We found the most misconceptions and instances of incomplete reasoning in Topics 3 and 4.

In contrast, Topic 6 responses include "global," "guess," "change," and "don't." There were three types of examples of Topic 6 responses. One response was to repeat the claim related to temperature change without adding additional information. For example, "Temperature will increase." Another type of response was to nominally cite global warming, as in "Why this is happening is because of global warming." Responses of the final type were unrelated to the science in question and included admissions regarding their actions or answers, including "I'm just guessing," "I don't know," and "I am correct." Despite their differences, all three types of Topic 6 responses were grouped together because they did not include particular mention of sea-ice data or scientific reasoning to explain how decreases in sea ice cause increases in temperature.

Step 4: Evaluations of LDA for explanations

To test how many occurrences of the words with each topic designation had characteristics uniquely associated with the topic, we analyzed LDA-generated probabilities for each explanation response. LDA computed the probability distribution of each explanation argument in each of the six topics. The overall mean of the probabilities of each explanation across the six topics was 16.6%. The mean of the highest explanation probability for each topic was 20.6%. We then used (16.6% + 20.6%)/2 = 18.65% as a threshold to reflect the specific topical characteristics. The accuracy indicator was the percentage of the explanations on a given topic that had probabilities above 18.65%. Across the six LDA-identified topics, the range of the accuracy-indicator value was 72.3% to 94.9% with a mean of 78.6%. The accuracy in topic assignment for explanations was considered in the scale of good to excellent (Harish et al. 2010). We further conducted qualitative analysis called topic intrusion to examine the validity of the derived topics in Step 4. Two doctoral students conducted this analysis together. Two hundred responses were randomly selected from the data corpus to examine the alignment of discovered topics with the students' written responses. Results showed a decent a topic accuracy as 84%, which reflects how much the discovered topics match the student response data.

In summary, the LDA results identified important words and unique words in each explanation topic, and these words were in turn used to interpret each topic in terms of the established claim–evidence–reasoning framework.

- Did not include data or reasoning (Topic 6)
- Cited data with or without a wrong claim (Topic 5)
- Connected data to a correct claim without providing reasoning (Topic 2)
- Connected data to a claim through reasoning based on ice being cold (Topic 4)
- Connected data to a claim through scientific but alternative reasoning based on the evaporation of water (Topic 3)
- Included scientific reasoning addressing the albedo effect (Topic 1)

Step 5: Differences in students' claims across the six identified explanation topics

We also examined how the presence of these topics in students' explanations related to students making correct claims. Figure 5 shows the mean percentage for each topic of the responses associated with a correct claim. As expected, Topic 1, which indicated the most scientific reasoning related to the albedo effect, was linked to the highest percentage of correct claims. The lowest percentage of correct claims was associated with Topic 5, indicating that data citation was not always associated with correct claims. These differences in distribution were statistically significant: $\chi^2(5) = 97.60$, $p < .001$. To determine whether two topics differed significantly in predicting correct claims, we conducted follow-up pair-wise chi-square comparisons. As Table 2 shows, of the 15 full comparisons, 10 were significant at $p = .05$. Differences between topics 2, 3, and 4 were not significant in part because the contents of these topics are very similar, as Fig. 4a shows. In sum, the LDA method identified explanation patterns that differed significantly in their associations with correct and incorrect claims.

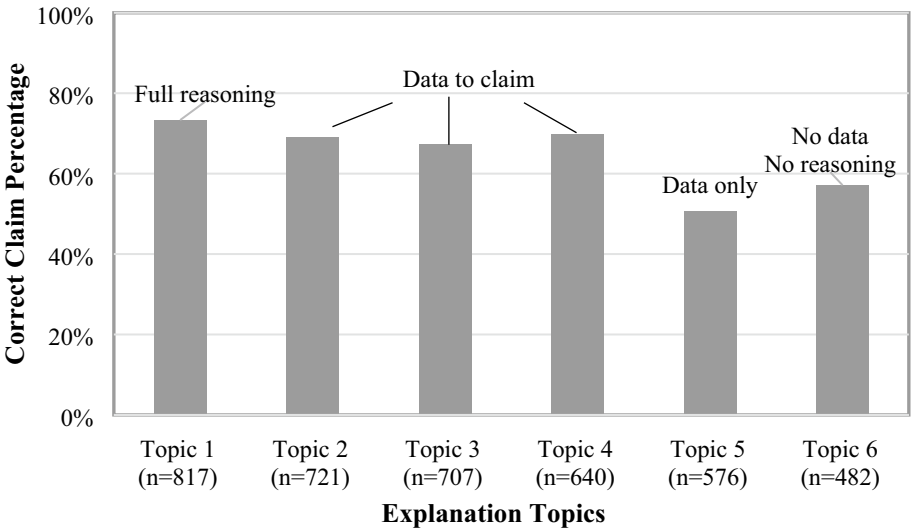


Fig. 5 The percentages of students who chose a correct claim across the six topics of explanation response

Table 2 Post hoc pair-wise comparisons for the explanations

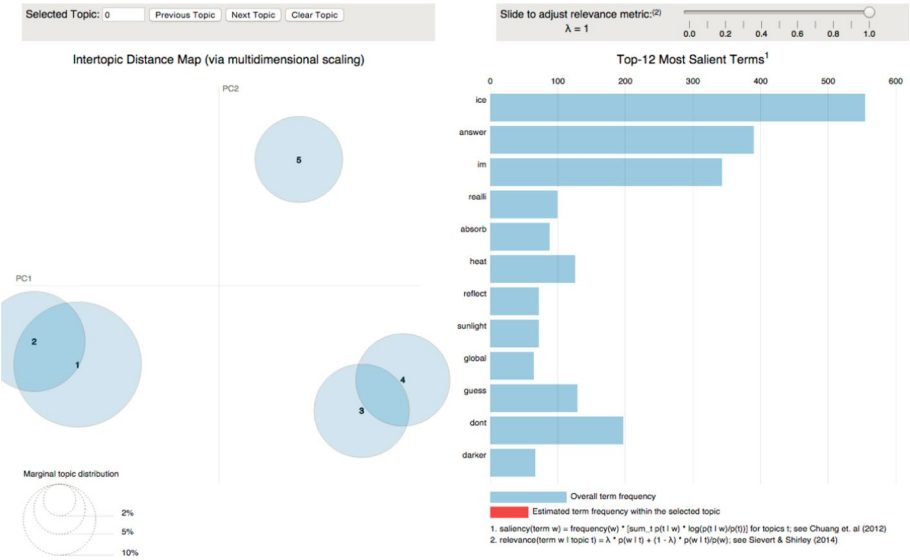
Comparison		χ^2	p
Topic 1	Topic 2	3.385	.065
Topic 1	Topic 3	7.773	.020*
Topic 1	Topic 4	2.226	.136
Topic 1	Topic 5	67.545	.000*
Topic 1	Topic 6	39.138	.000*
Topic 2	Topic 3	1.600	.449
Topic 2	Topic 4	0.073	.787
Topic 2	Topic 5	40.17	.000*
Topic 2	Topic 6	18.816	.000*
Topic 3	Topic 4	2.112	.348
Topic 3	Topic 5	32.213	.000*
Topic 3	Topic 6	14.004	.000*
Topic 4	Topic 5	43.606	.000*
Topic 4	Topic 6	21.391	.000*
Topic 5	Topic 6	4.202	.040*

*$p < .05$

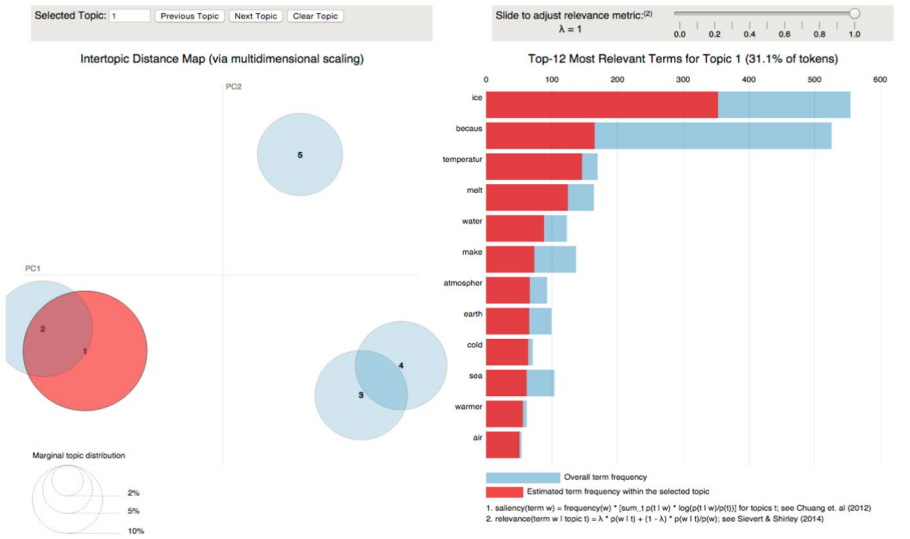
Step 3: Research question 2 on uncertainty attributions

Visualizing the LDA results

As Fig. 6 shows, LDA identified five topics. There are three content clusters. Topics 1 and 2 overlap, as do topics 3 and 4. Topic 5 is located far from the other four topics.



(**a**) The five topics identified in the uncertainty-rationale responses



(**b**) Topic 1 is highlighted

Fig. 6 A visualization of the LDA results for the uncertainty attributions. **a** On the left are the relative positions and sizes of the five topic groups (K = 5), and on the right are the top 12 most relevant words across all five topics. The size of the topic represents its prevalence, and the distances between the topics reflect dissimilarities between them. **b** Topic 1 highlighted in red with the list of most important words belonging to Topic 1 on the right

This map reveals similarities in content between topics 1 and 2 and between topics 3 and 4. The content of Topic 5 differs from those of the other four.

Characterizing the identified topics using the most representative topical words

By comparing the lists of important words shown in Table 3, we recognized that topics 1 and 2 include scientific words relevant to the albedo effect (Topic 2) and data on the melting of sea ice or on the consequences of warmer atmospheric temperatures (Topic 1). Topics 3 and 4 include words related to students' evaluations and explanations of their answers. Topics 3 and 4 share a number of words, including "I'm," "because," "question," "answer," and "guess." While they are similar, topics 3 and 4 differ slightly because Topic 3 focuses on negative self-evaluation—including words like "don't," "didn't," and "wasn't"—and Topic 4 focuses on positive self-evaluation—including words like "correct," "common," and "sense." Topic 5 responses can be characterized as sources of uncertainty rooted in previous beliefs (e.g. "It sounds like global warming, but global warming is not true"), experience (e.g. "I remember learning about this in previous science classes"), and knowledge (e.g. "This question reminded me of global warming"). Only one student response included all three topic clusters:

> I am very certain *[Topic 4, positive evaluation of their answer]* because we have talked about all of the effects in class *[Topic 5, personal, prior experience]* and also because the ice is one of the things that keeps the temperatures stable along with several other factors. *[Topic 1, science-related, unelaborated reasoning].*

Step 4: Evaluations of LDA for uncertainty

Using a similar procedure to that used for the explanations, we also tested for the uncertainty arguments how many of the instances in each topic presented the characteristics. The accuracy range was 72.3% to 94.9% with a mean of 78.6%, indicating very good topic assignment in terms identifying the correct characteristics. We also conducted a similar qualitative topic intrusion analysis here. Two hundred responses were randomly selected from the data corpus to examine the alignment of discovered topics with the students' written responses. Results showed a decent a topic accuracy as 78%, which reflects how much the discovered topics match the students' response data.

Step 5: Differences in the students' claims across the six identified explanation topics

We compared the average uncertainty ratings for each of the five topics (see Fig. 7). A one-way ANOVA indicated significant differences in the mean uncertainties for the five uncertainty-attribution topics: $F(4, 2417) = 19.74$, $p < .001$. For instance, when the students used scientific words to explain their uncertainty ratings—as they did in topics 1 and 2—they tended to provide significantly higher certainty ratings (indicating that they were more certain) than when they focused on self-evaluation—as they did in topics 3 and 4. It is obvious that the mean certainty rating was lowest when the students focused on negative self-evaluation. The students' personal sources of uncertainty—such as their prior experiences, knowledge, and skills—were associated with higher certainty ratings than were their

Table 3 Uncertainty attributions: words are presented in order of salience for each of the five topics

	Number of occurrences (N = 2472)	Most important words	Examples	Characterization
Topic 1	645	Ice, becaus, temperature*, melt*, water*, make, atmosphere*, earth*, cold*, sea, warmer*, air*	• What I know about the temperature of ice and how it effects the oceans • It must be getting hotter if the ice is melting • The ocean is usually cold so it might not be attracting heat • The reason is that hot air rises will could are cools. so when if the ice is melting then that means that there is only hot air left and the air is not cooling	Scientific source but not fully elaborated
Topic 2	557	Ice, heat*, absorb*, because, reflect*, sunlight*, darker*, light*, sea, surface*, color*, sun*	• Well the reading influenced my certainty because it says a darker surface absorbs more light making temperature • The question stated that darker surfaces absorb more sunlight than lighter surfaces. It's known that ocean water (which is usually a dark blue) is darker than ice (which is white)	Scientific source, fully elaborated
Topic 3	493	I'm, because, don't, answer, really*, question, guess, understand*, very, didn't*, happen, wasn't*	• I didn't really understand the question • I really don't know • I just took a guess in the first question • Because I kinda know that was the answer • Because I used my brain	Negative or unsure self-evaluation of the claim
Topic 4	491	Answer, I'm, because, correct*, question, guess, sense*, common, make, pretty, wrong*, predict	• My answer is a logical guess • It was an educated guess • I'm positive about my answer • I'm very confident • It all makes common sense	Positive self-evaluation of the claim

Table 3 (continued)

	Number of occurrences (N = 2472)	Most important words	Examples	Characterization
Topic 5	310	Because, learn*, global*, class*, answer, warm, inform, previous, influence, knowledge, warming*, science*	• I'm living in it and its what is happening • I remember learning about something like this in a previous science class and it makes sense • This is information that I don't remember very well from middle school • Global warming	Personal sources: prior beliefs, experience, or knowledge

*Most words exclusive to the topic are marked with an asterisk

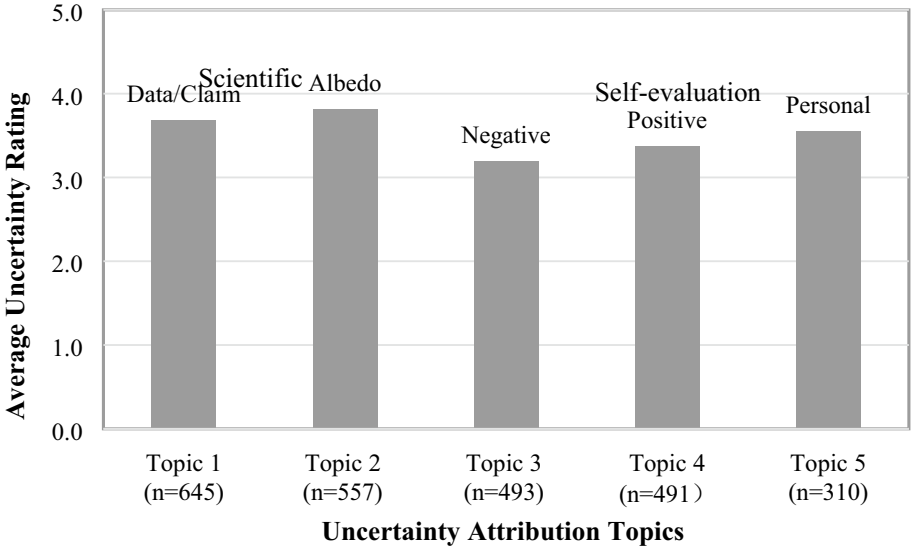


Fig. 7 The mean certainty ratings for all five uncertainty-attribution topics

self-evaluations. There were no apparent differences in the uncertainty ratings provided when students cited scientific and personal sources of uncertainty. We then conducted post-hoc analyses to identify significant differences between the mean uncertainty ratings of pairs of topics. As Table 4 shows, Tukey's tests indicated that 7 out of 10 pairs had statistically significant differences. Student responses that contained self-evaluation topics (Topic 3 and Topic 4) were associated with significantly lower uncertainty ratings than were those that contained scientific attribution sources, such as data/claim and albedo. Within the self-evaluation topics, responses that contained negative self-evaluations (Topic 3) had significantly lower uncertainty ratings than did responses that contained positive self-evaluations (Topic 4). Overall, responses containing negative self-evaluation generated significantly lower uncertainty ratings than did all of the other types of responses.

Table 4 Post-hoc pair-wise comparisons of uncertainty ratings

Comparison		*Mean difference*	*p*
Topic 1 (Data/Claim)	Topic 2 (Albedo)	−0.172	.310
Topic 1 (Data/Claim)	Topic 3 (Negative)	0.494	.000*
Topic 1 (Data/Claim)	Topic 4 (Positive)	0.268	.006*
Topic 1 (Data/Claim)	Topic 5 (Personal)	0.164	.202
Topic 2 (Albedo)	Topic 3 (Negative)	0.666	.000*
Topic 2 (Albedo)	Topic 4 (Positive)	0.440	.000*
Topic 2 (Albedo)	Topic 5 (Personal)	0.336	.001*
Topic 3 (Negative)	Topic 4 (Positive)	−0.226	.019*
Topic 3 (Negative)	Topic 5 (Personal)	−0.330	.000*
Topic 4 (Positive)	Topic 5 (Personal)	−0.104	.657

*$p < .01$

Discussion

Unsupervised text mining can automatically analyze the contents of written texts, identifying semantic patterns that can be interpreted using theoretical frameworks available to researchers. In this study, LDA identified patterns hidden in students' written texts, allowing the application of theoretical frameworks that were adapted to the study of students' uncertainty-infused scientific argumentation. LDA not only used salient words to identify various explanation and uncertainty-attribution types, but it also identified salient words representative of these types, enabling researchers to make interpretations according to the claim–evidence–reasoning framework (McNeill et al. 2006) and the uncertainty-attribution taxonomy (Lee et al. 2017). Unsupervised text mining is an innovative way to identify patterns in large-scale data. It can be particularly useful in developing scoring rubrics because topical methods of text mining like LDA provides teachers and instructional designers effective means by which they can identify all of the possible ways in which students might respond to an open-ended item.

In this study, LDA discovered meaningful patterns in students' explanations of their claims and in how these explanation patterns related to correct claims. Six explanation topics were identified. Even though the students were expected to fully elaborate scientific reasoning, using the data to justify their claims, only 817 of the 2472 explanations in which Topic 1 was identified contained fully elaborated scientific reasoning. The rest of the students used the opportunity for explanation to restate their claims, to cite data without providing reasoning, or to describe reasoning based on misconceptions or alternative ideas. Indeed, the students who provided scientific reasoning related to the albedo effect were significantly more likely to choose a correct claim than were those who did not. Students who only cited the data chose correct claims at significantly lower rates than did others. This suggests that if a student chooses a correct claim, this does not guarantee that the student can identify salient data and use this data to explain their claim scientifically (Osborne et al. 2004).

In addition, LDA showed the potential to discover more granular patterns than can qualitative coding in processing uncertainty-attribution responses and to better determine how these patterns influence certainty levels. The LDA identified three types of topics among the uncertainty attributions: "self-evaluation," "personal sources of uncertainty," and "scientific sources of uncertainty." The literature on uncertainty acknowledges that uncertainty can be attributed to subjective as well as objective sources (Allchin 2012). The LDA results reflect this duality and indicate that personal sources of uncertainty can be divided into: (1) evaluations of the current work and (2) evaluations of one's personal knowledge, experience, or skill set. When students evaluate their personal knowledge, experience, or skill set, their certainty ratings are higher than when they do not. This finding mirrors the theoretical work of Kahneman et al. (1982), which argued that people's expressions of uncertainty point either to the external world (in seeking more objective criteria) or to the state of their personal knowledge (due to internal ignorance). External attributions are directed either to frequencies of occurrence expressed by data or to causal mechanisms that can explain the occurrence. In this study, external attributions included finding scientific sources of uncertainty by examining specific data and instances of specific reasoning based on relevant scientific theory.

The LDA used in this study and its interactive visualization can also help teachers and instructional designers to support students' learning. For instance, based on the LDA results and the visualizations, teachers can quickly overview how students generally

constructed their scientific arguments and what the most popular patterns are. Also, teachers can even gain deeper insights into how students conduct scientific argumentation beyond the traditional claims, evidence and rebuttal. The LDA can also generate granular patterns which teachers can use to diagnostic students learning and provide more detailed feedback. Instructional designers can use the granular patterns to develop more comprehensive activities, scaffolds, and/or assessment frameworks to evaluate students.

The LDA used in this study has its limitations, however. The LDA and many other text mining algorithms only analyze the final product of the scientific argumentation (the text) and do not consider process information related to students' development of scientific-argumentation abilities over time. Students' development of scientific argumentation is a complex process, in which they connect their prior knowledge, their experience, and their understanding of the content to make sense of scientific phenomena. This process involves both cognitive and epistemological aspects (Sandoval 2003) that LDA and text mining are currently unable to capture. Moreover, as an unsupervised method of machine learning, LDA can only be used to survey text at the population level to find general topics. In addition, LDA is currently implemented in a high vocabulary specificity of a learning context, scientific argumentation. It may not work well which have more diverse cultural contexts and high diversity of vocabulary.

Another limitation concerns the unigram text model that underlies the LDA itself: LDA does not consider the respective positions of the words in a text document. LDA models arguments like "sea ice decreases increase temperature" and "ice decreases increase sea temperature" in the same way because LDA considers only single, unconnected words. This limitation can be overcome by using N-grams in addition one gram. An additional limitation of LDA is its topic composition and the situation for synonyms and meonyms: the same words can be found in multiple topics (as is revealed by our analysis of the topics discovered for the "explanation" and "uncertainty-attribution" responses) or words with similar meaning in different student responses. Unlike in a principal component analysis (decomposed basis), topics generated by LDA can overlap and are not always mutually independent and orthogonal. There are more structured approaches to addressing issues of topic composition, however, including hierarchical LDA and structural LDA. Topics produced by these algorithms can be joined together in hierarchical or nested structures, respectively, in which each node represents a topic distribution. In this way, topics can be made more obvious and distinct than they are in LDA.

LDA is just one of the many text mining techniques available. Further explorations of text mining tools could include information extraction, the answering of questions, and topic tracking. While LDA is conducted after students finish their argumentation tasks, topic tracking can enable real-time analyses of students' argumentation data, from which can be extracted information useful in providing feedback. For this reason, topic tracking could support the development of real time feedback mechanisms for students.

Conclusion

This study introduced a particular text mining technology—Latent Dirichlet Allocation (LDA)—that can be used to analyze the content of short, domain-specific, written scientific-argumentation responses. LDA was used to identify underlying semeiotic patterns among students' scientific-argumentation data. A dynamic visualization tool was used to facilitate the interpretation of the patterns discovered in light of established theories

concerning the learning of scientific argumentation. When it is applied and interpreted meaningfully, automatic text mining can significantly augment human pattern recognition and can be used as an effective survey tool. New insights and knowledge gained from LDA have the potential to transform teachers' practices through literature dissemination and through software applications embedded in learning environments.

Acknowledgements This work is supported by the National Science Foundation (NSF) of the United States under grant numbers DRL-1220756, and DRL-1418019. Any opinions, findings, and conclusions or recommendations expressed in this paper, however, are those of the authors and do not necessarily reflect the views of the NSF.

Compliance with ethical standards

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

Appendix

In the scientific argumentation data about the albedo effect described above, each open-ended response a student generates is a text document. That is, each explanation response can include several topics. So does each uncertainty attribution response. Students' explanation or uncertainty attribution responses are made up of topics that are made up of words. Therefore, LDA describes a document as a probability distribution of a mixture of topics, each of which is expressed with another probability distribution of words. Topics generated by LDA are a combination of words that contribute to the particular topic based on probabilities. LDA analysis results should be further interpreted by human insights about the context in which documents are generated. In this study, LDA is implemented in the following steps:

Step 1: Pre-processing and data preparation

To remove noise in the text data and format the data for input, pre-processing techniques are applied as follows:

1) All non-letter symbols, numbers, and punctuation are removed.
2) Common stop words such as "a," "and," "it," and "the" are removed.
3) Stemming is performed on the text to convert variations of the same word to a non-changing root word form. For instance, the root word "produc" captures several variations of the word like produced, producing, production, etc.
4) Infrequent words are filtered out.

These steps result in a data corpus in the form of a bag-of-words that takes into account the occurrences of words, but does not consider their ordering. Each text document is represented by a document matrix defined as a vector of the words found in the entire corpus along with the frequency of each word found in the document. This is the input for the LDA algorithm.

Step 2: Determining the number of topics K

For the LDA algorithm to work, the number of topics to be extracted from the data corpus, *K*, needs to be specified. *K* can be determined by statistical derivations or informed by researchers' insights about the documents. An optimum value of K can be determined through the Bayesian model selection and approximated using a harmonic mean estimator. The log-likelihood plots can show the best K value for the text corpus. Note that this statistically derived K value is not the absolute measure for K. Expert judgment based on data, knowledge, and experience can be important in selecting the most meaningful K value. In this study, we combine the log-likelihood method with human judgment to determine the K value for the scientific argument data corpus. Different K values were explored before determining the optimal number. Given the K value, LDA generates a list of relevant words for each topic (topical words) and which topics are contained in each document.

Step 3: Application of the LDA algorithm

Collapsed Gibbs sampling is applied as follows:

1) Each word in the corpus is randomly assigned to the K number of topics. Each topic now constitutes an initial random word distribution based on Dirichlet, which will be iteratively improved in the following steps.
2) For each word in a document,
- Compute the proportion of words assigned to a topic in the document, P(topic|document), and the proportion of words assigned to that topic from all documents, P(word|topic).
- Reassign the word to a new topic with the probability of P(topic|document) * P(word|topic).
3) Repeat Step 2 numerous times until the topic-word assignments are stabilized.
4) Use the topic assignments to calculate the proportion of topics in each document.

Distinctive words that appear in a topic and do not appear in other topics can be very useful to characterize the topic. If all the documents contain similar words, it is harder to cluster the words into topics, requiring expert evaluation.

References

Abdous, M., & He, W. (2011). Using text mining to uncover students' technology-related problems in live video streaming. *British Journal of Educational Technology, 40*(5), 40–49.

Abdous, M. H., Wu, H., & Yen, C. J. (2012). Using data mining for predicting relationships between online question theme and final grade. *Journal of Educational Technology & Society*, *15*(3), 77–88.

Allchin, D. (2012). Teaching the nature of science through scientific errors. *Science Education, 96*(5), 904–926.

Akçapınar, G. (2015). How automated feedback through text mining changes plagiaristic behavior in online assignments. *Computers & Education, 87*, 123–130.

Aufschnaiter, C. V., Erduran, S., Osborne, J., & Simon, S. (2008). Arguing to learn and learning to argue: Case studies of how students' argumentation relates to their scientific knowledge. *Journal of Research in Science Teaching, 45*, 101–131.

Beggrow, E. P., Ha, M., Nehm, R. H., Pearl, D., & 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*(1), 160–182.

Berland, L. K., & Reiser, B. J. (2009). Making sense of argumentation and explanation. *Science Education, 93*, 26–55.

Bricker, L. A., & Bell, P. (2008). Conceptualizations of argumentation from science studies and the learning sciences and their implications for the practices of science education. *Science Education, 92*, 473–493.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems* (pp. 288–296). New York: Curran Associates.

Chen, B. (2014). Visualizing semantic space of online discourse: The Knowledge Forum case. In *Proceedings of the 4th international conference on Learning Analytics and Knowledge (LAK '14), 24–28 March 2014, Indianapolis, IN* (pp. 271–272). New York: ACM. https://doi.org/10.1145/2567574.2567595.

Chen, B., Chen, X., & Xing, W. (2015). "Twitter archeology" of learning analytics and knowledge conferences. In *Proceedings of the 5th international conference on Learning Analytics and Knowledge (LAK '15), 16–20 March 2015, Poughkeepsie, NY* (pp. 340–349). New York: ACM. https://doi.org/10.1145/2723576.2723584.

Chen, N. S., Wei, C. W., & Chen, H. J. (2008). Mining e-Learning domain concept map from academic articles. *Computers & Education, 50*(3), 1009–1021.

Clark, D. B., & Sampson, V. (2008). Assessing dialogic argumentation in online environments to relate structure, grounds, and conceptual quality. *Journal of Research in Science Teaching, 45*, 293–321.

de Vries, E., Lund, K., & Baker, M. (2002). Computer-mediated epistemic dialogue: Explanation and argumentation as vehicles for understanding scientific notions. *The Journal of the Learning Sciences, 11*(1), 63–103.

Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment*, *5*(1), 1–36. Retrieved from https://ejournals.bc.edu/index.php/jtla/article/view/1640.

Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention. *Learning and Instruction, 22*(4), 271–280.

Duschl, R. A., Schweingruber, H. A., & Shouse, A. W. (2007). *Taking science to school: Learning and teaching science in grades K-8*. Washington, DC: National Academy Press.

Erduran, S., Simon, S., & Osborne, J. (2004). TAPping into argumentation: Developments in the application of Toulmin's argument pattern for studying science discourse. *Science Education, 88*, 915–933.

Ezen-Can, A., Boyer, K. E., Kellogg, S., & Booth, S. (2015). Unsupervised modeling for understanding MOOC discussion forums: A learning analytics approach. *Proceedings of the fifth international conference on learning analytics and knowledge* (pp. 146–150). New York: ACM. https://doi.org/10.1145/2723576.2723589.

Feldman, R. D. (1995). *Knowledge discovery in textual databases (KDT)*. Paper presented at the first international conference on knowledge discovery and data mining (KDD-95), August 20–21, Montreal, Canada.

Harish, B. S., Guru, D. S., & Manjunath, S. (2010). Representation and classification of text documents: A brief review. *International Journal of Computer Applications, Special Issue on RTIPPR, 2*(1), 110–119.

Hotho, A., Nürnberger, A., & Paaß, G. (2005). A brief survey of text mining. *Ldv Forum, 20*(1), 19–62.

Huff, J. D., & Nietfeld, J. L. (2009). Using strategy instruction and confidence judgments to improve metacognitive monitoring. *Metacognition and Learning, 4*(2), 161–176.

Hung, J. (2012). Trends of e-learning research from 2000 to 2008: Use of text mining and bibliometrics. *British Journal of Educational Technology, 43*(1), 5–16.

Janasik, N., Honkela, T., & Bruun, H. (2009). Text mining in qualitative research. *Organizational Research Methods, 12*(3), 436–460.

Kahneman, D., Slovic, S. P., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge: Cambridge University Press.

Kuhn, D. (1993). Science as argument: Implications for teaching and learning scientific thinking. *Science Education, 77*(3), 319–337.

Lee, H. S., Liu, O. L., Pallant, A., Roohr, K. C., Pryputniewicz, S., & Buck, Z. E. (2014). Assessment of uncertainty-infused scientific argumentation. *Journal of Research in Science Teaching, 51*(5), 581–605.

Lee, H. S., Pallant, A., Pryputniewicz, S., & Lord, T. (2017). Articulating uncertainty attribution as part of critical epistemic practice of scientific argumentation. In *Proceedings from CSCL 2017: Making*

a difference: Prioritizing equity and access in CSCL, 12th international conference on Computer Supported Collaborative Learning (CSCL) 2017 (Vol. 1, pp. 135–142). Philadelphia, PA: International Society of the Learning Sciences.

Lin, F. R., Hsieh, L. S., & Chuang, F. T. (2009). Discovering genres of online discussion threads via text mining. *Computers & Education, 52*(2), 481–495.

Liu, O. L., Rios, J. A., Heilman, M., Gerard, L., & Linn, M. C. (2016). Validation of automated scoring of science assessments. *Journal of Research in Science Teaching, 53*(2), 215–233.

McNeill, K. L., Lizotte, D. J., Krajcik, J., & Marx, R. W. (2006). Supporting students' construction of scientific explanations by fading scaffolds in instructional materials. *The Journal of the Learning Sciences*, *15*(2), 153–191.

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: National Academies Press.

NGSS Lead States. (2013). *Next generation science standards: For states, by states*. Washington, DC: National Academies Press.

Osborne, J., Erduran, S., & Simon, S. (2004). Enhancing the quality of argumentation in school science. *Journal of Research in Science Teaching, 41*(10), 994–1020.

Rosé, C., Wang, Y. C., Cui, Y., Arguello, J., Stegmann, K., Weinberger, A., et al. (2008). Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning. *International Journal of Computer-Supported Collaborative Learning*, *3*(3), 237–271.

Sadler, T. D., & Fowler, S. R. (2006). A threshold model of content knowledge transfer for socioscientific argumentation. *Science Education, 90*, 986–1004.

Sampson, V., & Clark, D. B. (2008). Assessment of the ways students generate arguments in science education: Current perspectives and recommendations for future directions. *Science Education, 92*, 447–472.

Sandoval, W. A. (2003). Conceptual and epistemic aspects of students' scientific explanations. *The Journal of the Learning Sciences, 12*(1), 5–51.

Sandoval, W. A., & Millwood, K. A. (2005). The quality of students' use of evidence in written scientific explanations. *Cognition and Instruction, 23*(1), 23–55.

Shermis, M. D., & Burstein, J. C. (Eds.). (2003). *Automated essay scoring: A cross-disciplinary perspective*. New Jersey: Routledge.

Sievert, C., & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. *Proceedings of the workshop on interactive language learning, visualization, and interfaces* (pp. 63–70).

Simosi, M. (2003). Using Toulmin's framework for the analysis of everyday argumentation: Some methodological considerations. *Argumentation, 17*, 185–202.

Southavilay, V., Yacef, K., Reimann, P., & Calvo, R. A. (2013). Analysis of collaborative writing processes using revision maps and probabilistic topic models. *Proceedings of the 3rd international conference on Learning Analytics and Knowledge (LAK '13), 8–12 April 2013, Leuven, Belgium* (pp. 38–47). New York: ACM. https://doi.org/10.1145/2460296.2460307.

Staley, K. W. (2014). Experimental knowledge in the face of theoretical error. In M. Boumans, G. Hon, & A. C. Petersen (Eds.), *Error and uncertainty in scientific practice: History and philosophy of technoscience* (pp. 39–56). London: Routledge.

Stone, N. J. (2000). Exploring the relationship between calibration and self-regulated learning. *Educational Psychology Review, 12*(4), 437–475.

Tane, J., Schmitz, C., & Stumme, G. (2004). Semantic resource management for the web: An e-learning application. *Proceedings of the WWW conference New York, USA, 2004* (pp. 1–10)

Tawfik, A. A., Law, V., Ge, X., Xing, W., & Kim, K. (2018). The effect of sustained vs. faded scaffolding on students' argumentation in ill-structured problem solving. *Computers in Human Behavior, 87*, 436–449.

Toulmin, S. (1958). *The uses of argument*. New York: Cambridge University Press.

Tseng, Y. H., Chang, C. Y., Rundgren, S. N. C., & Rundgren, C. J. (2010). Mining concept maps from news stories for measuring civic scientific literacy in media. *Computers & Education, 55*(1), 165–177.

Yu, C. H., Jannasch-Pennell, A., & Digangi, S. (2011). The qualitative report compatibility between text mining and qualitative research in the perspectives of grounded theory, content analysis, and reliability recommended APA citation compatibility between text mining and qualitative research in the perspect. *The Qualitative Report, 16*(3), 730–744.

Walton, D., Reed, C., & Macagno, F. (2008). *Argumentation schemes*. New York: Cambridge University Press.

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

Xing, W., Popov, V., Zhu, G., Horwitz, P., & McIntyre, C. (2019a). The effects of transformative and non-transformative discourse on individual performance in collaborative-inquiry learning. *Computers in Human Behavior, 98*, 267–276.

Xing, W., Tang, H., & Pei, B. (2019b). Beyond positive and negative emotions: Looking into the role of achievement emotions in discussion forums of MOOCs. *The Internet and Higher Education, 43*, 100690.

Zhang, F., & Litman, D. (2015). Annotation and classification of argumentative writing revisions. *Proceedings of the tenth workshop on innovative use of NLP for building educational applications* (pp. 133–143).

Zhu, G., Xing, W., Costa, S., Scardamalia, M., & Pei, B. (2019). Exploring emotional and cognitive dynamics of knowledge building in grades 1 and 2. *User Modeling and User-Adapted Interaction, 29*(4), 789–820.

Wanli Xing is an Assistant Professor of Educational Technology at University of Florida. His research interests are artificial intelligence, learning analytics, STEM education and online learning.

Hee-Sun Lee is a senior research scientist at the Concord Consortium. a non-profit educational technology research and development organization. Hee-Sun Lee is currently Co-PI on two NSF-funded projects, High-Adventure Science: Earth Systems and Sustainability and Investigating How to Enhance Scientific Argumentation through Automated Feedback in the Context of Two High School Earth Science Curriculum Units. Her main research areas include technology-enhanced, inquiry-based science curriculum design and evaluation, construct modeling, instrument design, science assessment validation and learning analytics. She earned her Ph.D. in science education from the University of Michigan, Ann Arbor, under the direction of Dr. Nancy Songer and was a postdoctoral scholar in Dr. Marcia Linn's Technology Enhanced Learning in Science (TELS) Center at the University of California, Berkeley.

Antonette Shibani is a lecturer at the University of Technology, Sydney. She was previously working as a Research Associate at the National Institute of Education, Nanyang Technological University, Singapore. Her research interests include Learning Analytics, particularly Text and Writing Analytics, Machine Learning and Learning Technologies.