# Data-Mining Textual Responses to Uncover Misconception Patterns

Joshua J. Michalenko
Rice University
jjm7@rice.edu

Andrew S. Lan
Princeton University
andrew.lan@princeton.edu

Andrew E. Waters
Rice University
aew2@rice.edu

Phillip J. Grimaldi
Rice University
phillip.grimaldi@rice.edu

Richard G. Baraniuk
Rice University
richb@rice.edu

## ABSTRACT

An important, yet largely unstudied problem in student data analysis is to detect *misconceptions* from students' responses to *open-response* questions. Misconception detection enables instructors to deliver more targeted feedback on the misconceptions exhibited by many students in their class, thus improving the quality of instruction. In this paper, we propose a new natural language processing-based framework to detect the common misconceptions among students' textual responses to short-answer questions. We propose a probabilistic model for students' textual responses involving misconceptions and experimentally validate it on a real-world student-response dataset. Experimental results show that our proposed framework excels at classifying whether a response exhibits one or more misconceptions. More importantly, it can also automatically detect the common misconceptions exhibited across responses from multiple students to multiple questions; this property is especially important at large scale, since instructors will no longer need to manually specify all possible misconceptions that students might exhibit.

## Keywords

Learning analytics, Markov chain Monte Carlo, misconception detection, natural language processing

## 1. INTRODUCTION

The rapid developments of large-scale learning platforms (e.g., MOOCs (edx.org, coursera.org) and OpenStax Tutor (openstaxtutor.org)) have enabled not only access to high-quality learning resources to a large number of students, but also the collection of student data at very large scale. The scale of this data presents a great opportunity to revolutionize education by using machine learning algorithms to *automatically* deliver personalized analytics and feedback to students and instructors in order to improve the quality of teaching and learning.

### 1.1 Detecting misconceptions from data

The predominant form of student data, their *responses* to assessment questions, contains rich information on their knowledge. Analyzing why a student answers a question incorrectly is of crucial importance to deliver timely and effective feedback. Among the possible causes for a student to answer a question incorrectly, exhibiting one or more *misconceptions* is critical, since upon detection of a misconception, it is very important to provide targeted feedback to a student to correct their misconception in a timely manner. Examples of using misconceptions to improve teaching include incorporating misconceptions to design better distractors for multiple-choice questions [10], implementing a dialogue-based tutor to detect misconceptions and provide corresponding feedback to help students self-practice, preparing prospective instructors by examining the causes of common misconceptions among students [19], and incorporating misconceptions into item response theory (IRT) for learning analytics [18].

The conventional way of leveraging misconceptions is to rely on a set of pre-defined misconceptions provided by domain experts [10, 19]. However, this approach is not scalable, since it requires a large amount of human effort and is domain-specific. With the large scale of student data at our disposal, a more scalable approach is to automatically detect misconceptions from data.

Recently, researchers have developed approaches for data-driven misconception detection; most of these approaches analyze students' response to *multiple-choice* questions. Examples of these approaches include detecting misconceptions in multiple-choice mathematics questions and modeling students' progress in correcting them [9] via the additive factor model [3], and clustering students' responses across a number of multiple-choice physics questions [20]. However, multiple-choice questions have been shown to be inferior to open-response questions in terms of pedagogical value [8]. Indeed, students' responses to open-response questions can offer deeper insights into their knowledge state.

To date, detecting misconceptions from students' responses to open-response questions has largely remained an unexplored problem. A few recent developments work exclusively with *structured* responses, e.g., sketches [17], short mathematical expressions [11], group discussions in a chemistry class [16], and algebra with simple syntax [4].

### 1.2 Contributions

In this paper, we propose a natural language processing framework that detects students' common misconceptions from their *textual* responses to open-response, short-answer questions. This problem is very difficult, since the responses are, in general, *unstructured*.

Our proposed framework consists of the following steps. First, we transform students' textual responses to a number of short-answer questions into low-dimensional textual feature vectors using several well-known word-vector embeddings. These tools include the popular Word2Vec embedding [12], the GLOVE embedding [15], and an embedding based on the long-short term memory (LSTM) neural network [6]. We then propose a new statistical model that jointly models both the transformed response textual feature vectors and expert labels on whether a response exhibits one or more misconceptions; these labels identify only *whether or not* a response exhibits one or more misconceptions but not *which* misconception it exhibits.

Our model uses a series of latent variables: the feature vectors corresponding to the correct response to each question, the feature vectors corresponding to each misconception, the tendency of each student to exhibit each misconception, and the confusion level of each question on each misconception. We develop a Markov chain Monte Carlo (MCMC) algorithm for parameter inference under the proposed statistical model. We experimentally validate the proposed framework on a real-world educational dataset collected from high school classes on AP biology.

Our experimental results show that the proposed framework excels at classifying whether a response exhibits one or more misconceptions compared to standard classification algorithms and significantly outperforms a baseline random forest classifier. We also compare the prediction performance across all three embeddings. More importantly, we show examples of common misconceptions detected from our dataset and discuss how this information can be used to deliver targeted feedback to help students correct their misconceptions.

## 2. DATASET AND PRE-PROCESSING

In this section, we first detail our short-answer response dataset, and then detail our pre-processing approach to convert responses into vectors using word-to-vector embeddings.

### 2.1 Dataset

Our dataset consists of students' textual responses to short-answer questions in high school classes on AP Biology administered on OpenStax Tutor [14]. Every response was labeled by an expert grader as to whether it exhibited one or more misconceptions. A total of $N = 386$ students each responded to a subset of a total of $Q = 1668$ questions; each response was manually labeled by one or multiple expert graders, resulting in a total of $\sim 60,000$ labeled responses. Since there is no clear rubric defining what is a misconception, graders might not necessarily agree on what label to assign to each response. Therefore, we trim the dataset to only keep responses that are labeled by multiple graders and they also assigned the same label, resulting in $13,099$ responses. We also further trim the dataset by filtering out students who respond to less than 5 questions and questions with less than 5 responses in every dataset. This subset contains $6,152$ responses.

| | $N$ | $Q$ | Sparsity (%) |
|---|---|---|---|
| Unit 1 | 47 | 77 | 0.280 |
| Unit 2 | 101 | 104 | 0.243 |
| Unit 3 | 73 | 91 | 0.236 |
| Unit 4 | 43 | 75 | 0.315 |

**Table 1: Dataset statistics.**

The questions in our dataset are drawn from the OpenStax AP biology textbook; we divide the full dataset into smaller subsets corresponding to each of the first four units [13], since different units correspond to entirely different sub-areas in biology. These units cover the following topics: Unit 1—The Chemistry of Life, Chapters 1-3, Unit 2—The Cell, Chapters 4-10, Unit 3—Genetics, Chapters 11-17, and Unit 4—Evolutionary Processes, Chapters 18-20. To summarize, we show the dimensions of the subsets of the data corresponding to each unit in Table 1. Since not every student was assigned to every question, the dataset is sparsely populated; Table 1 also shows the portion of responses that are observed in the trimmed data subsets, denoted as "sparsity".

### 2.2 Response embeddings

We first perform a pre-processing step by transforming each textual student response into a corresponding real-valued vector via three different word-vector embeddings. Our first embedding uses the Word2Vec embedding [12] trained on the OpenStax Biology textbook (an approach also mentioned in [2]), to learn embeddings that put more emphasis on the technical vocabulary specific to each subject. We create the feature vector for each response by mapping each individual word in the response to its corresponding feature vector, and then adding them together. Concretely, denote $\mathbf{x}_{i,j} = \{w_1, w_2, ..., w_{T_{i,j}}\}$ as the collection of words in the textual response of student $j$ to question $i$, where $T_{i,j}$ denotes the total number of words in this response (excluding common stopwords). We then map each word $w_t$ to its corresponding $D$-dimensional feature vector $r(w_t) \in \mathbb{R}^D$ using the trained Word2Vec model. We use $D = 10$ for the Word2Vec embedding. We then compute the student response feature vector as $\mathbf{f}_{i,j} = \sum_{t=1}^{T_{i,j}} r(w_t)$.

Our second word-vector embedding is a pre-trained GLOVE embedding with $D = 25$ [15]. The GLOVE embedding is very similar to the Word2Vec embedding, with the main difference being that it takes corpus-level word co-occurrence statistics into account. Moreover, the quality of the GLOVE embedding for common words is likely higher since it is pre-trained on a huge corpus (comparing to only the OpenStax Biology textbook for Word2Vec).

Both the Word2Vec embedding and the GLOVE embedding do not take word ordering into account, and for misconception classification, this drawback can lead to problems. For example, responses "If X then Y" and "If Y then X" may have completely different meanings depending on the context, where it's possible for one to exhibit a common misconception while the other one does not. Using the Word2Vec and GLOVE embeddings, these responses will be embedded to the same feature vector $\mathbf{f}_{i,j}$, making them indistinguishable from each other. Therefore, our third word-vector embedding is based on the long short-term memory (LSTM) neural network, which is a recurrent neural network that excels at capturing long-term dependencies in sequential data. Therefore, it can take word ordering into account, a feature that we believe is critical for misconception detection. We implement a 2-layer LSTM network with 10 hidden units and train it on the OpenStax Biology textbook. For each student

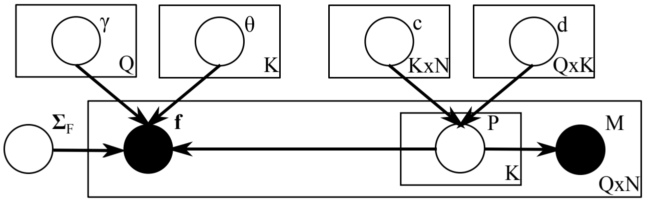


**Figure 1: Visualization of the statistical model. Black nodes denote observed data; white nodes denote latent variables to be inferred.**

response, we use the text as character-by-character inputs to the LSTM network and use the last layer's hidden unit activation values (stacked in a $D = 10$ dimensional vector) as its textual feature $\mathbf{f}_{i,j}$.

## 3. STATISTICAL MODEL

We now detail our statistical model; its graphical model is visualized in Figure 1. Concretely, let there be a total of $N$ students, $Q$ questions, and $K$ misconceptions. Let $M_{i,j} \in \{0, 1\}$ denote the binary-valued misconception label on the response of student $j$ to question $i$ provided by an expert grader, with $j \in \{1, \ldots, N\}$ and $i \in \{1, \ldots, Q\}$, where 1 represents the presence of (one or more) misconceptions, and 0 represents no misconceptions.

We transform the raw text of student $j$'s response to question $i$ into a $D$-dimensional real-valued feature vector, denoted by $\mathbf{f}_{i,j} \in \mathbb{R}^D$, via a pre-processing step (detailed in the previous section). Let $\Omega \subseteq \{1, \ldots, Q\} \times \{1, \ldots, N\}$ denote the subset of student responses that are labeled, since every student only responds to a subset of the questions.

We denote the *tendency* of student $j$ to exhibit misconception $k$, with $k \in \{1, \ldots, K\}$ as $c_{k,j} \in \mathbb{R}$, and the *confusion level* of question $i$ on misconception $k$, as $d_{i,k} \in \mathbb{R}$. Then, let $P_{i,j,k} \in \{0, 1\}$ denote the binary-valued latent variable that represents whether student $j$ exhibits misconception $k$ in their response to question $i$, with 1 denoting that the misconception is present and 0 otherwise. We model $P_{i,j,k}$ as a Bernoulli random variable

$$p(P_{i,j,k} = 1) = \Phi(c_{k,j} + d_{i,k}), \quad (i, j) \in \Omega, \tag{1}$$

where $\Phi(x) = \int_{-\infty}^{x} \mathcal{N}(t; 0, 1)\mathrm{d}t$ denotes the inverse probit link function (the cumulative distribution function of the standard normal random variable). Given $P_{i,j,k}$ $\forall k$, we model the observed misconception label $M_{i,j}$ as

$$M_{i,j} = \begin{cases} 0 & \text{if } P_{i,j,k} = 0 \ \forall k, \\ 1 & \text{otherwise,} \end{cases} \qquad (i, j) \in \Omega. \tag{2}$$

In words, a response is labeled as having a misconception if one or more misconceptions is present (given by the latent misconception exhibition variables $P_{i,j,k}$). Given $P_{i,j,k}$ $\forall k$, the textual response feature vector that corresponds to student $j$'s response to question $i$, $\mathbf{f}_{i,j}$, is modeled as

$$\mathbf{f}_{i,j} \sim \mathcal{N}(\boldsymbol{\gamma}_i + \sum_k P_{i,j,k}\boldsymbol{\theta}_k, \boldsymbol{\Sigma}_F), \quad \forall (i, j) \in \Omega, \tag{3}$$

where $\boldsymbol{\gamma}_i$ denotes the feature vector that corresponds to the correct response to question $i$, $\boldsymbol{\theta}_k$ denotes the feature vector that corresponds to misconception $k$, and $\boldsymbol{\Sigma}_F$ denotes the covariance matrix of the multivariate normal distribution characterizing the feature vectors. In other words, the feature vector of each response is a *mixture* of the feature vectors corresponding to the correct response to the question and each misconception the student exhibits. In the next section, we develop an MCMC inference algorithm to infer the values of the latent variables $\boldsymbol{\gamma}_i$, $\boldsymbol{\theta}_k$, $\boldsymbol{\Sigma}_F$, $P_{i,j,k}$, $c_{k,j}$, and $d_{i,k}$, given observed data $\mathbf{f}_{i,j}$ and $M_{i,j}$.

## 4. PARAMETER INFERENCE

We use a Gibbs sampling algorithm [5] for parameter inference under the proposed statistical model. The prior distributions of the latent variables are listed as follows:

$$\boldsymbol{\gamma}_i \sim \mathcal{N}(\boldsymbol{\mu}_\gamma, \boldsymbol{\Sigma}_\gamma), \boldsymbol{\theta}_k \sim \mathcal{N}(\boldsymbol{\mu}_\theta, \boldsymbol{\Sigma}_\theta), \boldsymbol{\Sigma}_F \sim IW(h_F, \mathbf{V}_F),$$
$$c_{k,j} \sim \mathcal{N}(\mu_c, \sigma_c^2), d_{i,k} \sim \mathcal{N}(\mu_d, \sigma_d^2),$$

where $IW(\cdot)$ denotes the inverse-Wishart distribution and $\boldsymbol{\mu}_\gamma$, $\boldsymbol{\Sigma}_\gamma$, $\boldsymbol{\mu}_\theta$, $\boldsymbol{\Sigma}_\theta$, $h_F$, $\mathbf{V}_F$, $\mu_c$, $\sigma_c^2$, $\mu_d$, and $\sigma_d^2$ are hyperparameters.

We start by randomly initializing the values of the latent variables $\boldsymbol{\gamma}_i$, $\boldsymbol{\theta}_k$, $\boldsymbol{\Sigma}_F$, $P_{i,j,k}$, $c_{k,j}$, $d_{i,k}$, $a_j$, and $\mu_i$ by sampling from their prior distributions. Then, in each iteration of our Gibbs sampling algorithm, we iteratively sample the value of each random variable from its full conditional posterior distribution. Specifically, in each iteration, we perform the following steps:

a) Sample $P_{i,j,k}$: We first sample the latent misconception indicator variable $P_{i,j,k}$ from its posterior distribution as

$$P_{i,j,k} = \begin{cases} 0 & \text{if } M_{i,j} = 0, \\ 1 & \text{if } M_{i,j} = 1 \text{ and } P_{i,j,k'} = 0 \ \forall \ k' \neq k, \\ \frac{r}{r+1} & \text{if } M_{i,j} = 1 \text{ and } \exists \ k' \neq k \text{ s.t. } P_{i,j,k'}{=}1, \end{cases}$$

where

$$r = \frac{p(\mathbf{f}_{i,j}|\boldsymbol{\gamma}_i, \boldsymbol{\theta}_k, \forall k, \boldsymbol{\Sigma}_F, P_{i,j,k'\neq k}, P_{i,j,k} = 1)}{p(\mathbf{f}_{i,j}|\boldsymbol{\gamma}_i, \boldsymbol{\theta}_k, \forall k, \boldsymbol{\Sigma}_F, P_{i,j,k'\neq k}, P_{i,j,k} = 0)} \cdot \frac{p(P_{i,j,k} = 1|c_{k,j}, d_{i,k})}{p(P_{i,j,k} = 0|c_{k,j}, d_{i,k})}.$$

Terms in these expressions are given by (1) and (3).

b) Sample $\boldsymbol{\gamma}_i$: We then sample the feature vector that corresponds to the correct response to each question, $\boldsymbol{\gamma}_i$, from its posterior distribution as $\boldsymbol{\gamma}_i \sim \mathcal{N}(\boldsymbol{\mu}_{\gamma_i}, \boldsymbol{\Sigma}_{\gamma_i})$ where

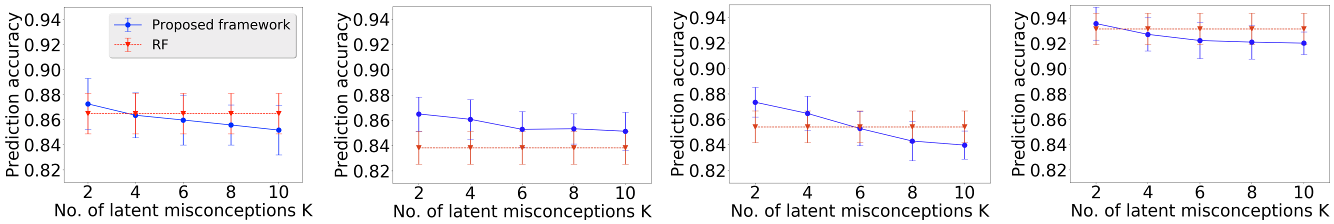
$$\boldsymbol{\mu}_{\gamma_i} = \boldsymbol{\Sigma}_{\gamma_i}\left(\boldsymbol{\Sigma}_\gamma^{-1}\boldsymbol{\mu}_\gamma + \boldsymbol{\Sigma}_F^{-1}\sum_{j:(i,j)\in\Omega}(\mathbf{f}_{i,j} - \sum_k P_{i,j,k}\boldsymbol{\theta}_k)\right),$$
$$\boldsymbol{\Sigma}_{\gamma_i} = (\boldsymbol{\Sigma}_\gamma^{-1} + n_i\boldsymbol{\Sigma}_F^{-1})^{-1},$$

where $n_i = \sum_j I\left((i, j) \in \Omega\right)$.

c) Sample $\boldsymbol{\theta}_k$: We then sample the feature vector that corresponds to each misconception, $\boldsymbol{\theta}_k$, from its posterior distribution as $\boldsymbol{\theta}_k \sim \mathcal{N}(\boldsymbol{\mu}_{\theta_k}, \boldsymbol{\Sigma}_{\theta_k})$ where

$$\boldsymbol{\mu}_{\theta_k} = \boldsymbol{\Sigma}_{\theta_k}\left(\boldsymbol{\Sigma}_\theta^{-1}\boldsymbol{\mu}_\theta + \boldsymbol{\Sigma}_F^{-1}\sum_{i,j:P_{i,j,k}=1}(\mathbf{f}_{i,j} - \boldsymbol{\gamma}_i - \sum_{k'\neq k} P_{i,j,k'}\boldsymbol{\theta}_{k'})\right),$$
$$\boldsymbol{\Sigma}_{\theta_k} = (\boldsymbol{\Sigma}_\theta^{-1} + n_k\boldsymbol{\Sigma}_F^{-1})^{-1},$$

where $n_k = \sum_{i,j} I\left(P_{i,j,k} = 1\right)$.



(a) Unit 1 (b) Unit 2 (c) Unit 3 (d) Unit 4

**Figure 2: Comparison of the prediction performance of the proposed model against RF on our AP Biology dataset using the ACC metric as the number of latent misconceptions $K$ varies, with the LSTM embedding.**

d) Sample $\boldsymbol{\Sigma}_F$: We then sample the covariance matrix $\boldsymbol{\Sigma}_F$ from its posterior distribution as

$$\boldsymbol{\Sigma}_F \sim IW\left(h_F + n, \mathbf{V}_F + \mathbf{M}\right),$$

where $n{=}\sum_{i,j} I\left((i,j) \in \Omega\right)$ and $\mathbf{M} = \sum_{i,j:(i,j)\in\Omega}(\mathbf{f}_{i,j} - \boldsymbol{\gamma}_i - \sum_k P_{i,j,k}\boldsymbol{\theta}_k)(\mathbf{f}_{i,j} - \boldsymbol{\gamma}_i - \sum_k P_{i,j,k}\boldsymbol{\theta}_k)^T$.

e) Sample $c_{k,j}$ and $d_{i,k}$: In order to sample $c_{k,j}$ and $d_{i,k}$, we first sample the value of the auxiliary variable $z_{i,j,k}$ (following the standard approach proposed in [1]) as

$$z_{i,j,k} \sim \mathcal{N}^{\pm}(c_{k,j} + d_{i,k}, 1), \forall (i,j) \in \Omega,$$

where $\mathcal{N}^{\pm}(\cdot)$ denotes the truncated normal random distribution truncated to the positive side when $P_{i,j,k} = 1$ and negative side when $P_{i,j,k} = 0$. We then sample $c_{k,j}$ from its posterior distribution as

$$c_{k,j} \sim \mathcal{N}(\mu_{c_{k,j}}, \sigma^2_{c_{k,j}}),$$

where $n_j = \sum_i I\left((i,j) \in \Omega\right)$, $\sigma^2_{c_{k,j}} = 1/(1/\sigma^2_c + n_j)$, and $\mu_{c_{k,j}}{=}\sigma^2_{c_{k,j}}\big(\mu_c/\sigma^2_c + \sum_{i:(i,j)\in\Omega}(z_{i,j,k} - d_{i,k})\big)$. We then sample $d_{i,k}$ from its posterior distribution as

$$d_{i,k} \sim \mathcal{N}(\mu_{d_{i,k}}, \sigma^2_{d_{i,k}}),$$

where $\sigma^2_{d_{i,k}}{=}1/(1/\sigma^2_d + n_i)$, and $\mu_{d_{i,k}} = \sigma^2_{d_{i,k}}(\mu_d/\sigma^2_d + \sum_{j:(i,j)\in\Omega}(z_{i,j,k} - c_{k,j}))$.

We run the iterations detailed above for a number of $T$ total iterations with a certain burn-in period, and use the samples of each latent variable to approximate their posterior distributions.

Parameter inference under our model suffers from the label-switching issue that is common in mixture models [5], meaning that the mixture components might be permuted between iterations. We employ a post-processing step to resolve this issue. We first calculate the augmented data likelihood at each iteration, (indexed by $\ell$) we then identify the iteration $\ell_{\max}$ with the largest augmented data likelihood, and permute the variables $\boldsymbol{\theta}^{\ell}_k$, $c^{\ell}_{k,j}$, and $d^{\ell}_{i,k}$ that best match the variables $\boldsymbol{\theta}^{\ell_{\max}}_k$, $c^{\ell_{\max}}_{k,j}$, and $d^{\ell_{\max}}_{i,k}$. After this post-processing step, we can simply calculate the posterior means of each one of these sets of variables by taking averages of their values across non burn-in iterations.

# 5. EXPERIMENTS

We experimentally validate the efficacy of the proposed framework using our AP Biology class dataset. We first compare the proposed framework against a baseline random forest (RF) classifier that classifies whether a student response exhibits one or more misconceptions. We then show common misconceptions detected in our datasets and discuss how the proposed framework can use this information to deliver meaningful targeted feedback to students that helps them correct their misconceptions.

## 5.1 Experimental setup

We run our experiments with $K \in \{2,4,6,8,10\}$ latent misconceptions with hyperparameters $\boldsymbol{\mu}_\gamma = \boldsymbol{\mu}_\theta = \mathbf{0}_D$, $\boldsymbol{\Sigma}_\gamma = \boldsymbol{\Sigma}_\gamma = \mathbf{V}_F = \mathbf{I}_D$, $h_F = 10$, $\mu_c = \mu_d = 0$, and $\sigma^2_c = \sigma^2_d = 1$, for a total of $T = 500$ iterations with the first 250 iterations as burn-in. We compare the proposed framework against a baseline random forest (RF) classifier[1] using the textual response feature vectors $\mathbf{f}_{i,j}$ to classify the binary-valued misconception label $M_{i,j}$, with 200 decision trees.

We randomly partition each dataset into 5 folds and use 4 folds as the training set and the other fold as the test set. We then train the proposed framework and RF on the training set and evaluate their performance on the test set, using two metrics: i) prediction accuracy (ACC), i.e., the portion of correct predictions, and ii) area under curve (AUC), i.e., the area under the receiver operating characteristic (ROC) curve of the resulting binary classifier [7]. Both metrics take values in $[0,1]$, with larger values corresponding to better prediction performance. We repeat our experiments for 20 random partitions of the folds.

For the proposed framework, the predictive probability that a response with its feature vector $\mathbf{f}_{i,j}$ exhibits a misconception, i.e., the probability that at least one of the $K$ latent misconception exhibition state variables take the value of 1, is given by $1 - \widehat{p}_{i,j}$, where

$$\begin{aligned}\widehat{p}_{i,j} &= p(M_{i,j} = 0 \,|\, \mathbf{f}_{i,j}, \boldsymbol{\gamma}_i, \boldsymbol{\Sigma}_F, \boldsymbol{\theta}_k, \forall k, c_{k,j}, d_{i,k})\\ &= \frac{p(\mathbf{f}_{i,j}|\boldsymbol{\theta}_k, P_{i,j,k} = 0, \forall k)\prod_k p(P_{i,j,k} = 0|c_{k,j}, d_{i,k})}{\sum_{P_{i,j,k},\,\forall k}(p(\mathbf{f}_{i,j}\,|\,\boldsymbol{\theta}_k, P_{i,j,k}\forall k)\prod_k p(P_{i,j,k}|c_{k,j}, d_{i,k}))},\end{aligned}$$

where in the last expression we omitted the conditional dependency of $\mathbf{f}_{i,j}$ on $\boldsymbol{\gamma}_i$ and $\boldsymbol{\Sigma}_F$ due to spatial constraints. For RF, the predictive probability is given by the fraction of decision trees that classifies $M_{i,j} = 1$ given $\mathbf{f}_{i,j}$.

## 5.2 Results and discussions

The number of latent misconceptions $K$ is an important parameter controlling the granularity of the misconceptions that

[1] The RF classifier achieves the best performance among a number of off-the-shelf baseline classifiers, e.g., logistic regression, support vector machines, etc. Therefore, we do not compare it against other baseline classifiers.

| | Unit 1 | | Unit 2 | | Unit 3 | | Unit 4 | |
|---|---|---|---|---|---|---|---|---|
| | ACC | AUC | ACC | AUC | ACC | AUC | ACC | AUC |
| Proposed framework | **0.789±0.014** | **0.762±0.027** | **0.774±0.015** | **0.758±0.023** | **0.779±0.019** | **0.752±0.020** | **0.887±0.011** | **0.774±0.029** |
| RF | 0.762±0.019 | 0.645±0.025 | 0.735±0.011 | 0.676±0.014 | 0.758±0.017 | 0.630±0.024 | 0.873±0.009 | 0.604±0.034 |
| Proposed framework | 0.867±0.014 | **0.762±0.048** | **0.870±0.010** | **0.821±0.024** | **0.893±0.017** | **0.794±0.039** | **0.953±0.015** | **0.892±0.047** |
| RF | **0.876±0.014** | 0.697±0.022 | 0.859±0.013 | 0.771±0.040 | 0.883±0.008 | 0.616±0.043 | 0.948±0.019 | 0.731±0.006 |
| Proposed framework | **0.873±0.042** | **0.772±0.093** | **0.865±0.025** | **0.829±0.044** | **0.873±0.027** | **0.792±0.061** | **0.936±0.032** | **0.832±0.094** |
| RF | 0.865±0.035 | 0.711±0.086 | 0.838±0.028 | 0.722±0.043 | 0.854±0.028 | 0.697±0.057 | 0.931±0.025 | 0.709±0.105 |

**Table 2: Performance comparison on misconception label classification of a textual response in terms of the prediction accuracy (ACC) and area under the receiver operating characteristic curve (AUC) of the proposed framework against a random forest (RF) classifier, using the AP Biology dataset and the Word2Vec (top), GLOVE (middle), and LSTM (bottom) embeddings.**

we aim to detect. Figure 2 shows the comparison between the proposed framework using different values of $K$ and RF using the ACC metric with the LSTM embedding. We see an obvious trend that, as $K$ increases, the prediction performance decreases. The likely cause of this trend is that the proposed framework tends to overfit as the number of latent misconceptions grows very large since some of our datasets do not contain very rich misconception types. Moreover, the number of common misconceptions varies across different units, with Unit 2 likely containing more misconception types than Units 1 and 4.

We then compare the performance of the proposed framework against RF on misconception label classification in Table 2 using $K = 2$ and all three embeddings. The proposed framework significantly outperforms RF (1–4% using the ACC metric and 4-18% using the AUC metric) on almost all 4 data subsets using every embedding. The only case where the proposed framework does not outperform RF is on Unit 1 using the GLOVE embedding. We postulate that the reason for this result is that this unit is about chemistry and has a lot of responses with more chemical molecular expressions than words; therefore, the proposed framework does not have enough textual information to exhibit its advantages (grouping responses that share the same misconceptions into clusters) over the RF classifier.

Both the proposed framework and RF perform much better using the GLOVE and LSTM embeddings than the Word2Vec embedding. This result is likely due to the fact that these embeddings are more advanced than the Word2Vec embedding: the GLOVE embedding considers additional word co-occurrence statistics than the Word2Vec embedding, is trained on a much larger corpus, and has a higher dimension $D = 25$, while the LSTM embedding is the only embedding that takes word ordering into account. Moreover, both algorithms perform best on Unit 4, which is likely due to two reasons: i) the Unit 4 subset has a larger portion of its responses labeled, and ii) Unit 4 is about evolution, which results in responses that are much longer and thus contains richer textual information.

## 5.3 Uncovering common misconceptions

We emphasize that, in addition to the proposed framework's significant improvement over RF in terms of misconception label classification, it features great interpretability since it identifies common misconceptions from data. As an illustrative example, the following responses from multiple students across two questions are identified to exhibit the same misconception in the Unit 4 subset using the Word2Vec embedding:

> *Question 1*: People who breed domesticated animals try to avoid inbreeding even though most domesticated animals are indiscriminate. Evaluate why this is a good practice.
> *Correct Response*: A breeder would not allow close relatives to mate, because inbreeding can bring together deleterious recessive mutations that can cause abnormalities and susceptibility to disease.
> **Student Response 1**: Inbreeding can cause a rise in unfavorable or detrimental traits such as genes that cause individuals to be prone to disease or have unfavorable mutations.
> **Student Response 2**: Interbreeding can lead to harmful mutations.

> *Question 2*: When closely related individuals mate with each other, or inbreed, the offspring are often not as fit as the offspring of two unrelated individuals. Why?
> *Correct Response*: Inbreeding can bring together rare, deleterious mutations that lead to harmful phenotypes.
> **Student Response 3**: Leads to more homozygous recessive genes thus leading to mutation or disease.
> **Student Response 4**: When related individuals mate it can lead to harmful mutations.

Although these responses are from different students to different questions, they exhibit one common misconception, that inbreeding leads to harmful mutations. Once this misconception is identified, course instructors can deliver the targeted feedback that inbreeding only brings together harmful mutations, leading to issues like abnormalities, rather than directly leading to harmful mutations.

Moreover, the proposed framework can automatically discover common misconceptions that students exhibit without input from domain experts, especially when the number of students and questions are very large. Specifically, in the example above, we are able to detect such a common misconception that 4 responses exhibit by analyzing the 1016 responses in the AP Biology Unit 4 dataset; however, it would not likely be detected if the number of responses was smaller and fewer students exhibited the misconception. This feature makes it an attractive data-driven aid to domain experts in designing educational content to address student misconceptions.

We show another example that the proposed framework can automatically group student responses to the same group according to the misconceptions they exhibit. The example shows two detected common misconceptions among students' responses to a single question in the Unit 2 subset using the LSTM embedding:

> *Question*: What is the primary energy source for cells?
> *Correct response*: Glucose.
> **Student responses with misconception** 1:
> a) sunlight b) sum c) The sun d) he sun?
> **Student responses with misconception** 2:
> a) ATP b) adenosine triphosphate
> c) ATPPPPPPPPPPPPPP d) atp mitochondria

We see that the proposed framework has successfully identified two common misconception groups, with incorrect responses that list "sun" and "ATP" as the primary energy source for cells. Note that the LSTM embedding enables the framework to assign the full and abbreviated form of the same entity ("adenosine triphosphate" and "ATP") into the same misconception cluster, without employing any preprocessing on the raw textual response data. The likely reason for this result is that our LSTM embedding is trained on a character-by-character level on the OpenStax Biology textbook, where these terms appear together frequently, thus enabling the LSTM to transform them into similar vectors. This observation highlights the importance of using good, information-preserving word-vector embeddings for the proposed framework to maximize its capability of detecting common misconceptions.

## 6. CONCLUSIONS AND FUTURE WORK

In this paper, we have proposed a natural language processing-based framework for detecting and classifying common misconceptions in students' textual responses. Our proposed framework first transforms their textual responses into low-dimensional feature vectors using three existing word-vector embedding techniques, and then estimates the feature vectors characterizing each misconception, among other latent variables, using a proposed mixture model that leverages information provided by expert human graders. Our experiments on a real-world educational dataset consisting of students' textual responses to short-answer questions showed that the proposed framework excels at classifying whether a response exhibits one or more misconceptions. Our proposed framework is also able to group responses with the same misconceptions into clusters, enabling the data-driven discovery of common misconceptions without input from domain experts. Possible avenues of future work include i) automatically generate the appropriate feedback to correct each misconception, ii) leverage additional information, such as the text of the correct response to each question, to further improve the performance on predicting misconception labels, iii) explore the relationship between the dimension of the word-vector embeddings and prediction performance, and iv) develop embeddings for other types of responses, e.g., mathematical expressions and chemical equations.

## 7. REFERENCES

[1] J. H. Albert and S. Chib. Bayesian analysis of binary and polychotomous response data. *J. Am. Stat. Assoc.*, 88(422):669–679, June 1993.

[2] S. Bhatnagar, M. Desmarais, N. Lasry, and E. S. Charles. Text classification of student self-explanations in college physics questions. In *Proc. 9th Intl. Conf. Educ. Data Min.*, pages 571–572, July 2016.

[3] H. Cen, K. R. Koedinger, and B. Junker. Learning factors analysis – A general method for cognitive model evaluation and improvement. In *Proc. 8th. Intl. Conf. Intell. Tutoring Syst.*, pages 164–175, June 2006.

[4] M. Elmadani, M. Mathews, A. Mitrovic, G. Biswas, L. H. Wong, and T. Hirashima. Data-driven misconception discovery in constraint-based intelligent tutoring systems. In *Proc. 20th Int. Conf. Comput. in Educ.*, pages 1–8, Nov. 2012.

[5] A. Gelman, J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin. *Bayesian Data Analysis*. CRC press, 2013.

[6] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1–32, Nov. 1997.

[7] H. Jin and C. X. Ling. Using AUC and accuracy in evaluating learning algorithms. *IEEE Trans. Knowl. Data Eng.*, 17(3):299–310, Mar. 2005.

[8] S. Kang, K. McDermott, and H. Roediger III. Test format and corrective feedback modify the effect of testing on long-term retention. *Eur. J. Cogn. Psychol.*, 19(4-5):528–558, July 2007.

[9] R. Liu, R. Patel, and K. R. Koedinger. Modeling common misconceptions in learning process data. In *Proc. 6th Intl. Conf. on Learn. Analyt. & Knowl.*, pages 369–377, Apr. 2016.

[10] J. K. Maass and P. I. Pavlik Jr. Modeling the influence of format and depth during effortful retrieval practice. In *Proc. 9th Intl. Conf. Educ. Data Min.*, pages 143–149, July 2016.

[11] T. McTavish and J. Larusson. Discovering and describing types of mathematical errors. In *Proc. 7th Intl. Conf. Educ. Data Min.*, pages 353–354, July 2014.

[12] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, Sep. 2013.

[13] OpenStax Biology. https://openstax.org/details/biology, 2016.

[14] OpenStax Tutor. https://openstaxtutor.org/, 2016.

[15] J. Pennington, R. Socher, and C. D. Manning. GloVe: Global vectors for word representation. In *Proc. ACM SIGDAT Conf. Emp. Method. Nat. Lang. Process.*, pages 1532–1543, Oct. 2014.

[16] H. J. Schmidt. Students' misconceptions—Looking for a pattern. *Sci. Educ.*, 81(2):123–135, Apr. 1997.

[17] A. Smith, E. N. Wiebe, B. W. Mott, and J. C. Lester. SketchMiner: Mining learner-generated science drawings with topological abstraction. In *Proc. 7th Intl. Conf. Educ. Data Min.*, pages 288–291, July 2014.

[18] K. K. Tatsuoka. Rule space: An approach for dealing with misconceptions based on item response theory. *J. Educ. Meas.*, 20(4):345–354, Dec. 1983.

[19] D. Tirosh. Enhancing prospective teachers' knowledge of children's conceptions: The case of division of fractions. *J. Res. Math. Educ.*, 31(1):5–25, Jan. 2000.

[20] G. Zheng, S. Kim, Y. Tan, and A. Galyardt. Soft clustering of physics misconceptions using a mixed membership model. In *Proc. 9th Intl. Conf. Educ. Data Min.*, pages 658–659, July 2016.