A Deep Transfer Learning Approach to Modeling Teacher Discourse in the Classroom

Emily Jensen
University of Colorado Boulder

Samuel L. Pugh
University of Colorado Boulder

Sidney K. D'Mello
University of Colorado Boulder

ABSTRACT

Teachers, like everyone else, need objective reliable feedback in order to improve their effectiveness. However, developing a system for automated teacher feedback entails many decisions regarding data collection procedures, automated analysis, and presentation of feedback for reflection. We address the latter two questions by comparing two different machine learning approaches to automatically model seven features of teacher discourse (e.g., use of questions, elaborated evaluations). We compared a traditional open-vocabulary approach using n-grams and Random Forest classifiers with a state-of-the-art deep transfer learning approach for natural language processing (BERT). We found a tradeoff between data quantity and accuracy, where deep models had an advantage on larger datasets, but not for smaller datasets, particularly for variables with low incidence rates. We also compared the models based on the level of feedback granularity: utterance-level (e.g., whether an utterance is a question or a statement), class session-level proportions by averaging across utterances (e.g., question incidence score of 48%), and session-level ordinal feedback based on pre-determined thresholds (e.g., question asking score is medium [vs. low or high]) and found that BERT generally provided more accurate feedback at all levels of granularity. Thus, BERT appears to be the most viable approach to providing automatic feedback on teacher discourse provided there is sufficient data to fine tune the model.

CCS CONCEPTS

• **Computing methodologies** → Machine learning; Modeling and simulation; • **Human-centered computing**;

KEYWORDS

Automated Feedback, Deep Learning, Natural Language Processing, Teacher Discourse, Teaching Analytics

ACM Reference Format:
Emily Jensen, Samuel L. Pugh, and Sidney K. D'Mello. 2021. A Deep Transfer Learning Approach to Modeling Teacher Discourse in the Classroom. In *LAK21: 11th International Learning Analytics and Knowledge Conference (LAK21), April 12–16, 2021, Irvine, CA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3448139.3448168

LAK21, April 12–16, 2021, Irvine, CA, USA

ACM ISBN 978-1-4503-8935-8/21/04…$15.00
https://doi.org/10.1145/3448139.3448168

1 INTRODUCTION

Teachers, like anyone, need feedback in order to improve their effectiveness in the classroom [6, 20, 26, 68]. Since they teach almost daily, teachers have plenty of opportunities to practice and refine these skills. However, isolated practice is not enough; achieving expertise might require deliberate practice, which often takes place under the guidance of a coach [28]. Such a coach would design training tasks at an appropriate level of difficulty and provide feedback and guidance to steer teachers towards continuous improvement [24, 26, 27].

Unfortunately, current professional development (PD) opportunities are a far cry from this form of deliberate practice. One issue is that conference-style PDs, large events with lectures on a wide range of topics, are largely ineffective [7, 11, 30, 31, 74] since they do not provide the individual support teachers need to apply the knowledge to their own classrooms. Another method of PD involves classroom observation, where peers or supervisors give feedback, ideally based on validated evaluation protocols and rubrics. Unfortunately, these methods are usually evaluative rather than formative and prohibitively time consuming to implement on a frequent basis, which is what teachers need to improve [4].

This gap in PD leaves an exciting opportunity for AI-driven learning analytics to empower teachers to guide their own growth. Automated approaches can provide teachers with feedback specific to their own practice without requiring cost-prohibitive human observation. Teachers can use the analytics and supporting tools to reflect on their practice, set goals, and track progress, either individually or with a peer or a coach. And because the feedback is computed automatically, it can be more objective and reliable than human judgments which might be affected by human error and biases.

Accordingly, we focus on the development of an automatic system that provides teachers with objective feedback on the quality of their discourse (or teacher talk) in authentic classrooms. This entails several design considerations. First, the system should be able to record data of sufficient quality for automated analysis (see [12, 21, 57]). Some factors that may influence data collection are the measured variables (which influence the type of data to collect), the difficulty or expense of collecting data (due to costly or intrusive sensors), security and privacy concerns, and autonomy of teachers to record their own data.

The next design consideration pertains to the algorithms for automatically analyzing the recorded data. While previous work has primarily used engineered features and supervised traditional classifiers to analyze teacher discourse (see discussion below), recent improvements in state-of-the-art natural language processing techniques (such as word embeddings and transformers [14, 46]) warrant investigation of how these deep learning methods can be applied to the analysis of teacher discourse. These newer methods also

require much more training data than the traditional approaches, which may be prohibitive in some applications.

Finally, insights from the analyzed data should be presented to the teacher in a manner that makes the resulting analytics actionable, for example, enabling teachers to reflect on their practice when preparing for future lessons. There is currently little consensus on how this information should be presented to teachers. For example, [57] presents an orchestration graph to teachers, which gives a detailed account of their class activity over the entire class session. Alternatively, [41] presents teachers with an overall score in each teaching outcome for a given class session, which gives teachers a more high-level understanding of their teaching for a given lesson.

Having addressed the first design consideration in previous work [21, 41], we now turn to the latter two considerations by comparing two methods for modeling teacher discourse. First, we consider a traditional open-vocabulary approach, which uses n-grams and Random Forest classifiers to provide automated feedback, and which has yielded the most accurate modeling results for this problem to date [41]. We compare this to a state-of-the-art natural language processing approach (BERT) that uses transfer learning (on large domain-independent corpora) and fine-tuning (on our domain-specific data). We first consider how these models compare for different sized datasets and then analyze them at different levels of feedback granularity.

1.1 Related Work

We review work on automated approaches for analyzing teacher discourse quality and design considerations when presenting teachers with feedback.

1.1.1 Automated Analysis of Teacher Discourse. We focus on automated analysis of teacher discourse, primarily through the use of recorded audio. Although recording classroom sessions for teacher assessment is not new [2, 16, 32], the transition to automatically analyzing the recordings has been relatively recent. Some lines of research have used classroom audio to identify general classroom activities (e.g., time spent in lecturing vs. discussion) using turn-taking dynamics [78] or by analyzing utterance timing, language, and acoustic features [23]. Other work has focused on identifying the amount of teacher versus student talk (as in the startup teachfx.com). More recent work has focused on identifying general discourse features such as the frequency of question asking and the types of questions [10, 19, 54, 69] and instructional talk (compared to classroom management) [69]. Additionally, a few studies have begun to focus on modeling specific discourse feature that extend beyond questions, such as restating student ideas [41, 71, 72].

Traditional methods of automated teacher discourse analysis generally rely on one of two methods. The first, most common, method uses feature engineering based on automatic speech recognition (ASR) transcripts. It involves computing high-level features, such as linguistic features that span word, sentence, and discourse levels, and using them as inputs to standard supervised classifiers that can detect the focal discourse features in a generalizable manner [43] . The second method uses an open-vocabulary approach, which uses the words (and short phrases comprised of two or three words) themselves as features rather than more abstract representations [19, 70]. With the exception of [71, 72] the studies mentioned above all employ one of these approaches.

Recent advances in natural language processing have introduced a potential new method of automated teacher analysis - deep transfer learning. Deep transfer learning methods leverage the massive amounts of available online text data and the power of artificial neural networks with multiple hidden layers to achieve state of the art performance on a range of natural language processing (NLP) tasks, including text classification, the task considered in this study. Specifically, the introduction of the transformer architecture [75] in 2017 sparked a wave of deep transfer learning models that have advanced the state of the art in NLP. Rather than using purely supervised learning (above two approaches), transfer (machine) learning takes a model trained on one dataset/task and adapts it for another [55]. This entails two steps: pre-training and fine-tuning. During pre-training, the transformer uses large amounts (e.g., gigabytes) of text to learn the contextual meaning of words using domain-independent tasks. The trained model serves as the starting point for subsequent fine-tuning where it is then augmented with an output layer specific to the current task and tuned (update the parameters) using small amounts of domain-specific data. Recently, [71, 72] have used deep learning methods to detect specific dialogic strategies in mathematics classrooms. However, these studies used human-transcribed (rather than automatically transcribed) utterances, so it is unclear how these models can address ASR errors, which will inevitably occur. These studies also did not ensure teacher-independent training folds, so overfitting is also a concern.

1.1.2 Presenting Feedback on Teaching. Teacher feedback systems generally serve one of two main purposes: (1) providing information on student learning (e.g., identifying at-risk students, real-time class orchestration); and (2) providing information on teaching pedagogy and effectiveness (e.g., improving professional development) [17, 52]. Most systems are deployed in hybrid or virtual learning environments [17] and take advantage of extensive log data in the form of interactions with course materials, social interactions, assessment results, and time spent engaged with the platform [77]. For teachers in particular, the most common data recorded generally measures student time on platform and engagement with discussion boards [65]. There is a growing field of multimodal learning analytics which seeks to collect and analyze data in a more traditional face-to-face classroom setting. For example, [57] uses five different sensors to identify types of classroom activities and later displays an overview of the class session for the teacher to view.

There is also considerable variation in how analytics are presented to teachers. Some systems aim to provide real-time feedback that teachers can immediately act upon during a class session [50]. For systems related to student learning, analytics often centers on information like student engagement [5] or if a student needs immediate assistance [3, 38]. Research focused on teacher pedagogy has used virtual simulation technology like Augmented Reality/Virtual Reality to allow practice before a live class [8, 47, 48]. Additional work has used synchronous feedback, often from peers or supervisors, as a form of coaching [39, 62]. Other platforms show feedback after a class session in the form of trends in various metrics over time, such as the EdSight project [1], which aims to promote teacher

reflection by providing feedback based on student surveys. However, this example and others often rely on self-reported perceptions of the lesson rather than objective feedback [57].

There are very few studies comparing the effectiveness of feedback design choices. In particular, it is important to consider the granularity of presented information and intended insights. In [76] the authors introduce a process model for teacher feedback which entails the following steps: awareness, reflection, sensemaking, and impact. For example, if a teacher does not understand the relative effectiveness of different classroom activities, providing them with a detailed breakdown of their class time spent on these activities will not give them insight for how to improve their teaching. Feedback is perceived as more useful when it contains more complete data in the form of more metrics and more visualizations [77]. However, this approach poses a risk because automated feedback is not always accurate and there is a chance of presenting misleading information. Extremely fine-grained feedback is also potentially risky because teachers often have difficulty using analytics to identify next steps for improving their practice [52, 64]. This is in line with the case study in [49], where teachers reported a desire of having interpretations along with raw data.

1.2 Contribution and Research Questions

We expand previous work by comparing open-vocabulary (previous work) and deep transfer learning methods (current study) for automatic teacher discourse classification from recorded teacher audio in authentic classrooms. Our data includes 16,977 automatically transcribed teacher utterances, expert-coded for seven discourse features such as asking questions, providing elaborated feedback, and specifying learning goals. For the standard approach, we train Random Forest (RF) models using an open-vocabulary approach focusing on n-grams [41]. For the deep transfer learning approach, we use state of the art natural language processing techniques by fine-tuning an existing Bidirectional Encoder Representations from Transformers (BERT [22]) model.

Because data collection resources vary, some applications will need to choose automated models that can provide accurate results with limited data. We then pose Research Question 1 (*RQ1*): What is the data-accuracy tradeoff of standard vs. deep learning approaches? We address this question by sampling different quantities of our training data and comparing the two approaches as a function of data quantity.

Beyond availability of data, feedback analytics may be presented at different levels of granularity depending on the application and on the accuracy of the underlying models. Accordingly, we analyze our results at three levels of granularity. First, we investigate how these two methods compare for classification of individual utterances. This level of feedback is the most fine-grained analysis available for teachers and entails tagging individual utterances with discourse labels and comparing them to human-codes (ground-truth labels). Next, we compute class-session-level proportions for each discourse variable by aggregating across utterances and correlating the computer-predicted with human coded proportions. This level of feedback provides an aggregated overview on the incidence of each discourse variable and is at an intermediate level of granularity. Finally, we discretize the session-level proportions into ordinal categories (low, medium, and high) based on percentile cutoffs from the entire corpus. This level of feedback gives teachers an understanding of their performance relative to their peers and is at the coarsest level of granularity.

Whereas utterance-level feedback requires a high degree of accuracy because the feedback is resolved at the level of individual utterances, the intermediate session-level proportional feedback can accommodate a modicum of prediction errors because it capitalizes on the power of aggregation to eliminate noise. The ordinal-level feedback takes this a step further by not providing any numeric feedback, instead focusing on relative performance and should be even more tolerant to errors. Thus, for our second research question, (*RQ2*) we ask how the two methods compare with respect to these three levels of granularity?

2 METHOD

2.1 Teacher Talk Data

We used data from a prior study, which is detailed in [41], and only report aspects germane to the present study.

We recruited 16 English Language Arts (ELA) teachers from three suburban school districts in Pennsylvania. These teachers were trained to independently record their own classroom talk. Each teacher recorded at least four sessions of two different classes. From these, we identified a total of 127 recordings (out of 142 original recordings) that were usable for automated analysis.

We automatically segmented and transcribed each recording using the IBM Watson speech recognizer [61], which achieved an average word error rate of 0.28. Of the total 35,142 utterances, we randomly selected 200 sequential utterances from each recording to be coded by raters trained and supervised by ELA content experts. The final dataset included 16,977 coded utterances, with an average reliability of 0.81 (Gwet's AC [35]).

The dataset includes codes for seven teacher discourse variables, which are drawn from the literature on teaching effectiveness, student engagement, and achievement. Specifically, we focus on dialogic discourse, which emphasizes taking students' ideas seriously [29] and increasing opportunities for deeper cognition and engagement [44, 53, 59]. Based on this framework, we distinguished between questions and statements and also coded questions based on whether they were authentic (open-ended) questions with no pre-specified response (see also [42, 51]). Additionally, we included discourse variables inspired by Shernoff [66, 67] and Grossman [34], which include goal specificity [29, 66], use of ELA-specific terms [25, 36], cognitive level [13, 33, 58, 73], and elaborated feedback. Finally, we distinguished instructional talk from other talk such as classroom management.

Descriptions, examples, and incidences of the discourse variables are in Table 1. Note that these categories are not mutually exclusive; for example, Authentic Questions are a specific type of Questions, so the percentages do not add up to 100. Although some of the selected discourse variables have low prevalence rates in the dataset, they were selected based on their documented or hypothesized influence on student achievement (e.g., [53]) and low incidence is not to be equated to low impact. The low incidence also presents an important challenge for automated methods which often struggle to learn with unbalanced datasets [40], resulting in directly modeling

Table 1: Description of key teacher discourse variables ordered from highest to lowest prevalence.

Discourse Variable	Definition	Prevalence	Positive Example
Instructional Talk	Focuses on the lesson and learning goals rather than on other topics, such as classroom management or procedural talk.	81%	Let's think about the tone of this poem.
Questions	Requests for information.	31%	Do you have a pencil?
Goal Specified	Extent to which the teacher explains the process and end goals of a particular activity.	9%	Your writing partner should give you three overall comments, before editing supporting details.
ELA Terms	The use of discipline-specific terms in teacher talk.	9%	Ensure that you include a topic sentence in each one of your paragraphs.
Elaborated Evaluation	Expression of judgment or correctness of a student's utterance with explicit guidance for student learning and thinking.	6%	That's right. You're dying with each breath, and this is what the poet tries to bring to the consciousness of the beloved.
Authentic Questions	Open-ended question for which the teacher does not have a pre-scripted answer.	5%	What was your reaction to the end of the story?
High Cognitive Level	Emphasizes analysis (e.g., compare, interpret, synthesize, etc.) rather than reports or recitation of facts (e.g., define, recall, identify)	4%	How were their reactions to the accident different?

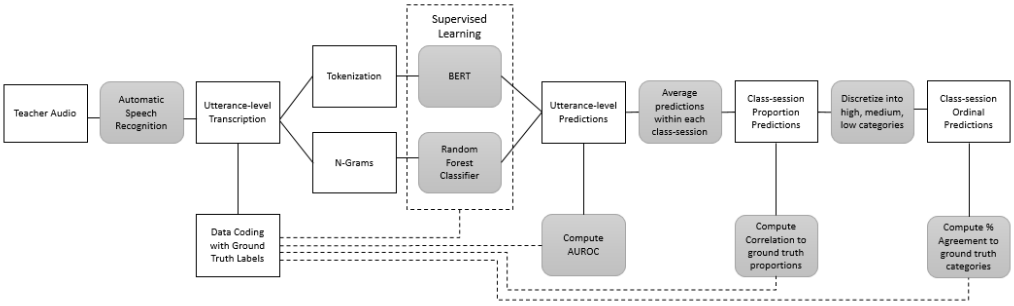


Figure 1: Overview of the automated analysis and feedback generation procedure.

the data at the proportion level in lieu of utterance-level modeling [43] .

2.2 Machine Learning Procedures

We adopted a supervised classification approach to predict the presence or absence of the discourse variables in each utterance. In particular, we compared two supervised classifiers: Random Forest Classifier (RF) and deep transfer learning using Bidirectional Encoder Representations from Transformers (BERT). Both RF and BERT output a prediction from 0 to 1 that an utterance reflects a given discourse variable, which was taken as the starting point for subsequent aggregation. The general approach is illustrated in Figure 1

2.2.1 Random Forest Classifier. We derived features for the Random Forest classifier using a bag of n-grams approach, which computes counts of words and phrases from the automatically transcribed utterances. We used unigrams (words), bigrams and trigrams (two- and three-word phrases) for our bag of n-gram features. Additionally, we filtered bigrams and trigrams using a pointwise mutual information [18] of 2, to ensure that meaningful n-grams ("topic sentence") were preserved, and not simply frequent words that occur together ("and then"). We also filtered the data to only include n-grams that occur with a minimum frequency in the corpus (we experimented with values of 1%, 2%, and 3%). We then trained Random Forest classifiers to predict the presence of the discourse variables in each utterance using the n-gram features described above. Separate binary classifiers were trained for each discourse variable (i.e., each model learns to predict the presence of only one of the variables, for example, whether an utterance is classified as an authentic question [1] or not [0]). We used the scikit-learn [56] library's implementation of the Random Forest Classifier with 100 estimators (the default).

2.2.2 BERT.. We used transfer learning to fine-tune BERT models to predict the presence of the discourse variables in each utterance. This entails starting with a BERT model pre-trained on large amounts of unlabeled data and fine-tuning it on our dataset of transcribed utterances and corresponding labels. Unlike the bag of

n-grams approach used for the Random Forest models, BERT processes the automatically transcribed utterances using WordPiece tokenization [63]. Here, an utterance is first split into a sequence of words, or parts of words. Each unique word or word piece is then converted to an integer according to the model's pre-specified vocabulary, and the sequence of integers is used as input to the model. As with the Random Forest models, a separate model was trained for each discourse variable. We started with the transformers [79] library's implementation of the BertForSequenceClassification model and the BertTokenizer and fine-tuned the BERT model for two epochs using a batch size of 32.

2.2.3 Cross Validation and Majority Sampling. We used random teacher-level nested 8-fold cross validation for both classifiers. This means that all the utterances for a given teacher were either included in the training set or the testing set, but never in both. This approach promotes generalizability to new teachers because it ensures a model is never trained and evaluated on utterances from the same teacher. Importantly, we used identical cross validation folds for the RF and BERT models to ensure that differences in performance are not an artifact of the folds used. Due to the imbalance of the discourse variables in our data, with several discourse variables having very low base rates, we used the imblearn [45] library to undersample the majority class during training of the RF models; distributions of the test set were unchanged.

3 RESULTS

3.1 (RQ1) Comparing Models Across Different Dataset Sizes

To investigate the tradeoff between data and model accuracy (RQ1), we randomly sampled 25%, 50%, and 75% of the utterances from the full dataset (16,977 utterances). We repeated the experiment for 10 iterations. Sampling was done without replacement within an iteration, but utterances could be repeated across iterations. We trained the RF and BERT models (7 discourse variables $\times$ 2 classifiers $\times$ 10 iterations) on the sampled data using the 8-fold cross validation procedure described above. We used the same sampled datasets and cross validation folds for equitable comparison across the two classifiers. We focused on utterance-level accuracy for this analysis since the other accuracy metrics are derived from these utterance-level predictions. We used the area under the receiver operating characteristic curve (AUROC) as our evaluation metric, which compares true positive and false positive rates across different classification thresholds. An AUROC of .5 represents chance performance.

Figure 2 shows the mean AUROC and 95% confidence interval across the 10 iterations for each of the sampling rates (25%, 50%, 75% and 100% [no sampling]). We used a bootstrap method to statistically compare AUROC values for the two models for each discourse variable and each iteration. This analysis was performed using the pROC package [60] in R with 2,000 bootstrap permutations. For each discourse variable, we adjusted the resulting p-values across the 10 iterations using a false discovery rate correction [9]. Sampling rates where one of the models performed significantly better (FDR corrected *ps* < .05) on the discourse variable in 7 or more of the 10 iterations are marked with an asterisk on the x-axis in Figure 2

Results varied by sampling rate. At the 25% sampling rate, BERT outperformed RF for Instructional Talk and ELA Terms, while RF outperformed BERT for Authentic Questions. There was no clear best model for the remaining four discourse variables. At the 50% sampling rate, BERT outperformed RF on all discourse variables except Authentic Questions, where RF had a significant advantage, and High Cognitive Level, for which there was no significant difference. Interestingly, the two discourse variables for which BERT did not outperform RF at either the 25% or 50% sampling rate (Authentic Questions and High Cognitive Level) had the lowest base rates of all variables examined (.05 and .04, respectively), which indicates that when using smaller amounts of data, RF may be better a better model for these variables. To this point, the BERT models' accuracy was at chance level (AUROC of 0.5) for the 25% sampling rate for Authenticity, whereas RF was above chance. At the 75% and 100% sampling rates, BERT significantly outperformed RF on all discourse variables except Authentic Questions, where there was no clear difference between the two models.

3.2 (RQ2) Comparing Models Across Different Levels of Granularity

Next, we compared the two models on the full dataset at the three levels of granularity – utterance-level, session-level proportions, and session-level ordinal categories (See Introduction).

3.2.1 Utterance-level results. Mean AUROC values for each discourse variable across the 10 iterations are reported in Table 2, along with 95% confidence intervals. We also used the bootstrap method to compare the AUROC values and adjusted the resulting p-values with a false discovery rate correction, as described in 2.4. We report the number of iterations with statistically significant (FDR corrected *ps* < .05) differences in Table 2. We found that BERT significantly outperformed RF for all 10 iterations on five of the discourse variables: Instructional Talk, Questions, Goal Specified, ELA Terms, and Elaborated Evaluation. It performed significantly better than RF for 9 iterations for High Cognitive Level. For authentic questions, BERT only outperformed RF for 6 of the iterations. There were no iterations where RF significantly outperformed BERT. Overall, the results strongly favor BERT vs. RF on the full dataset.

3.2.2 Session-level Proportion Results. We next compared the performance of RF and BERT models at the class session level by averaging the utterance-level ground-truth human codes and the RF/BERT predictions to the session level ($N = 127$). We then computed the Pearson correlation between the human- and computer-proportions. For each iteration, we used the Meng, Rosenthal, and Rubin's z test [37] for overlapping correlations to determine if there were significant differences among the two models. We again applied an FDR correction [9] to the resulting p-values to account for multiple testing across the 10 iterations (Table 3). BERT yielded significantly higher session-level correlations than RF for Instructional Talk, Goal Specified, and ELA Terms (all 10 iterations), Questions and Elaborated Evaluation (9 out of 10 iterations). Interestingly, BERT was only better than Authentic Questions for three iterations

Table 2: Utterance-level results for BERT and RF models, reported as mean AUROC across iterations. We also report the number of iterations where the difference in AUROC of the two models was statistically significant.

Discourse Variable	BERT	Random Forest	# Significant (out of 10)
Instructional Talk	0.828 [0.827-0.830]	0.762 [0.761-0.763]	10
Questions	0.830 [0.825-0.834]	0.762 [0.761-0.764]	10
Goal Specified	0.878 [0.875-0.881]	0.826 [0.824-0.828]	10
ELA Terms	0.895 [0.891-0.899]	0.763 [0.762-0.764]	10
Elaborated Evaluation	0.861 [0.858-0.863]	0.814 [0.812-0.816]	10
Authentic Questions	0.725 [0.711-0.739]	0.705 [0.701-0.710]	6
High Cognitive Level	0.868 [0.863-0.872]	0.850 [0.847-0.852]	9
Mean	0.841	0.783	

Table 3: Class session-level results for BERT and RF models, reported as mean Pearson *r* across iterations. We also report the number of iterations where the difference in correlations of the two models is statistically significant.

Discourse Variable	BERT	Random Forest	# Significant (out of 10)
Instructional Talk	0.545 [0.521-0.569]	0.262 [0.247-0.276]	10
Questions	0.694 [0.666-0.722]	0.529 [0.520-0.538]	9
Goal Specified	0.626 [0.614-0.639]	0.445 [0.434-0.456]	10
ELA Terms	0.695 [0.677-0.714]	0.292 [0.276-0.307]	10
Elaborated Evaluation	0.465 [0.444-0.487]	0.306 [0.295-0.317]	9
Authentic Questions	0.350 [0.249-0.452]	0.207 [0.170-0.244]	3
High Cognitive Level	0.526 [0.507-0.546]	0.438 [0.429-0.446]	0
Mean	.557	.354	

and there were no significant differences for the other seven iterations. Finally, the difference in correlations were not significant for any iterations of High Cognitive Level. Whereas there was no statistical advantage to using BERT over RF on these two low-prevalence variables, the magnitude of the correlations was higher for BERT for these two variables. Overall, the small utterance-level advantage of BERT over RF in AUROCs (mean of .841 vs. .783 across all seven variables) was compounded (mean correlation of .557 vs. 354) when utterances were aggregated to the session level.

3.2.3 Session-level Ordinal Results. We lastly compared the models after we discretized the class session-level proportions into high, medium, and low ordinal categories using the percentile splits pertaining to each distribution (RF, BERT, actual proportions) for each discourse variable. We considered two different splits: 33:67 and 15:85, which indicate the cutoff for the low and high categories, respectively (i.e., proportions < .33 are categorized as low; >.67% as high; median in-between). We chose these splits to examine the tradeoff between model accuracy and ordinal category size (i.e., the medium category contains 70% vs. 33% of instances for the 15:85 and 33:67 splits, respectively). Accuracy was computed as the diagonal agreement between the model assignments of category (low, medium, high) with ground-truth alignments at the observational level. The mean accuracy scores (and 95% CI) for each discourse variable across 10 iterations are shown in Table 4.

We statistically analyzed the data using mixed effects logistic regression models. Specifically, we regressed agreement (1 or 0) on model (RF [reference group] or BERT) with iteration and class session as (categorical) random intercepts. The resulting odds ratios are shown in Table 4 where values greater than 1 indicate an advantage of BERT vs. RF. We found that the BERT model consistently yielded higher agreement than the RF model for the 33:67 split with the exception of Authenticity, where the two models were tied. The differences were less pronounced for the 15:85 split, where BERT significantly outperformed RF for three of the discourse variables; the differences were marginally significant for two additional variables. Overall, as could be expected, agreement was higher for the 15:85 split (BERT average of 69%) than the 33:67 split (BERT average of 52%) because the former is less discriminating (i.e., the middle category contains 70% of the cases). This would explain why BERTs advantages over RF were more pronounced for the more discriminating 33:67 split.

4 DISCUSSION

4.1 Main Findings

We compared two machine learning approaches to model teacher discourse features with an eye for providing automated feedback for teacher learning. The first was a traditional open-vocabulary approach using a Random Forest model to predict the presence of key discourse variables in automatically transcribed teacher speech. We then compared this approach to BERT, a state-of-the-art natural language processing model which learns the contextual semantics of words from domain-independent training data, upon which the model is fine-tuned to the current domain of teacher talk.

Due to varying opportunities for data collection, our first task was to investigate the data-accuracy tradeoff between these two

Table 4: Percent Agreement [95% CI] of session-level ordinal feedback fir 15:85 and 33:67 splits across 10 iterations, along with Odds Ratio values (reference is RF) for each split.

	Percent Agreement [95% CI across iterations]				Odds Ratio (OR)	
	15:85 Split		33:67 Split		15:85 Split	33:67 Split
Discourse Variable	BERT	RF	BERT	RF		
Instructional Talk	0.691 [0.677-0.706]	0.645 [0.629-0.660]	0.511 [0.495-0.527]	0.394 [0.380-0.408]	1.67***	2.58***
Questions	0.698 [0.673-0.724]	0.686 [0.670-0.701]	0.584 [0.559-0.609]	0.498 [0.484-0.513]	1.17	2.58***
Goal Specified	0.680 [0.664-0.697]	0.604 [0.591-0.617]	0.526 [0.514-0.538]	0.490 [0.473-0.507]	2.09***	1.42**
ELA Terms	0.745 [0.722-0.767]	0.517 [0.512-0.533]	0.654 [0.636-0.673]	0.387 [0.369-0.405]	7.64***	9.97***
Elaborated Evaluation	0.656 [0.635-0.677]	0.634 [0.621-0.647]	0.513 [0.497-0.528]	0.446 [0.434-0.457]	1.261[1]	1.74***
Authentic Questions	0.656 [0.636-0.675]	0.626 [0.614-0.638]	0.417 [0.373-0.462]	0.433 [0.403-0.463]	1.201[1]	0.91
High Cognitive Level	0.706 [0.692-0.719]	0.688 [0.679-0.698]	0.435 [0.427-0.443]	0.403 [0.390-0.416]	1.22	1.29*
Mean	0.690	0.629	0.520	0.436	-	-

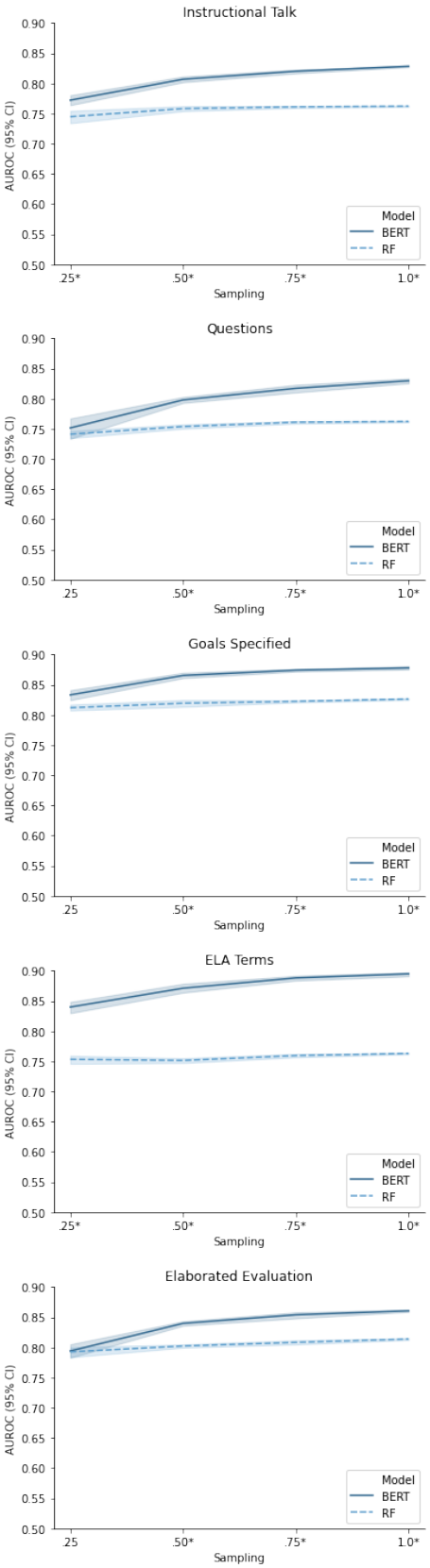
***p < .001; ** p < .01; * p < .05; [1] p < .057

models. Specifically, *RQ1* asked whether one model would be a better choice if the available training data were limited. We addressed this question through a sampling experiment where we trained each model using different sized partitions of our dataset. Perhaps unsurprisingly, we found that for both models, larger datasets generally yielded better model performance. Whereas RF had some advantageous for variables with low incidence rates when 25%-50% of the data was included, BERT generally outperformed RF for larger datasets. Compared to some NLP datasets which can contain millions of training samples, our own dataset was relatively modest at 16,977 samples. We hypothesize that BERT performance may improve even more using a larger dataset than is currently available to us.

Our next task was to consider how these two approaches compared when presenting data at different granularities (*RQ2*). We first considered utterance-level feedback, which identifies whether each utterance contains a given discourse variable or not. This type of feedback is the most specific form of feedback, which would allow teachers to identify positive and negative examples of behaviors they are trying to improve in the classroom. Compared to a traditional in-class observation by a peer or supervisor, utterance-level feedback is similar to the observer pointing out specific moments in class that the teacher excelled or needed improvement. We found BERT clearly outperformed the RF model for five of the seven discourse variables; the differences were negligible for the other two variables. Overall, BERT had a higher mean AUROC score of .841 compared to the RF model's mean AUROC of .783, but both easily outperformed chance (AUROC of 0.5). These results suggest that both models might be capable of providing feedback at this level of granularity. That said, it remains an important empirical question of how accurate these models must be in order to provide exemplar-based feedback to teachers because providing false positives as examples of particular utterances will erode trust.

We next considered session-level proportion feedback, which generates an overall score for each discourse variable in a given class session. This type of feedback provides a class-level summary per variable per class that teachers can directly focus on improving. By pooling across tens or even hundreds of utterances, it can mitigate utterance-level modeling errors. Additionally, teachers may be more capable of connecting this value with actionable goals for future lessons (e.g., increase Questions from 25% to 30%). Similar to the utterance-level feedback, we found the BERT clearly outperformed RF for five out of the seven discourse variables. Overall, BERT had an average correlation of .557 compared to .354 for RF, a larger relative improvement (57%) than the utterance-level AUROC scores (7.4% improvement).

Finally, we considered session-level ordinal feedback, which categorizes the scores from RQ2 into high, medium, or low categories relative to the other teachers in the dataset. This feedback provides another level of abstraction, which can hopefully protect against inevitable errors in the automated analysis. Feedback at this level may also serve as further motivation for teachers to improve since it reports their score relative to other teachers. However, this categorical feedback may be harder for teachers to interpret to make actionable insights because it is less clear what improvement looks like (e.g., moving up from 40th percentile to 60th percentile is still medium). We compared the two approaches using two different splits between categories. We found that BERT had generally higher agreement scores than RF, though the differences were larger and more consistent for the more discriminating 33:67 split compared



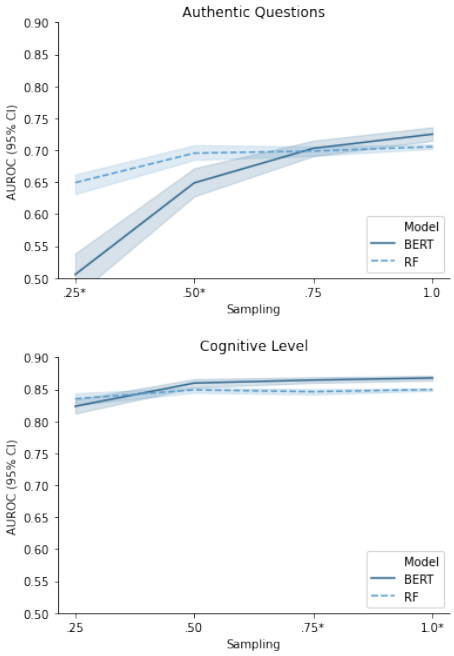


Figure 2: Mean (with 95% CI) utterance-level AUROCs of RF (dashed line) and BERT (solid line) across sampling rates of 25%, 50%, 75%, and 100%. Sampling rates where one of the models performed significantly better on 7 or more iterations are marked with a * on the x-axis.

to the 15:85 split. For both models, agreement was higher using the 15:85 split, where more instances are considered medium, which is perhaps less informative for teachers. Thus, BERT also appears to be superior to RF for the coarsest level of feedback granularity.

4.2 Limitations and Future Work

Our study was limited in a few aspects. First, our dataset was relatively homogeneous in that it contained data from only 16 teachers from a similar geographic region. Hence, our models may not generalize well to new dialects or other sources of teacher variation. More robust data collection is needed to further explore this possibility. Second, we only considered a limited number of machine learning models in our study. Specifically, we built off of previously verified methods and chose RF to exemplify the traditional open-vocabulary approach; we used BERT to exemplify the deep transfer learning approach to automated teacher feedback. As this area of natural language processing research continues to rapidly evolve, more models should be considered. Finally, our results apply specifically to discourse in English Language Arts classrooms; future work should similarly consider automated feedback methods in other subject areas to provide a holistic assessment of its value for teachers.

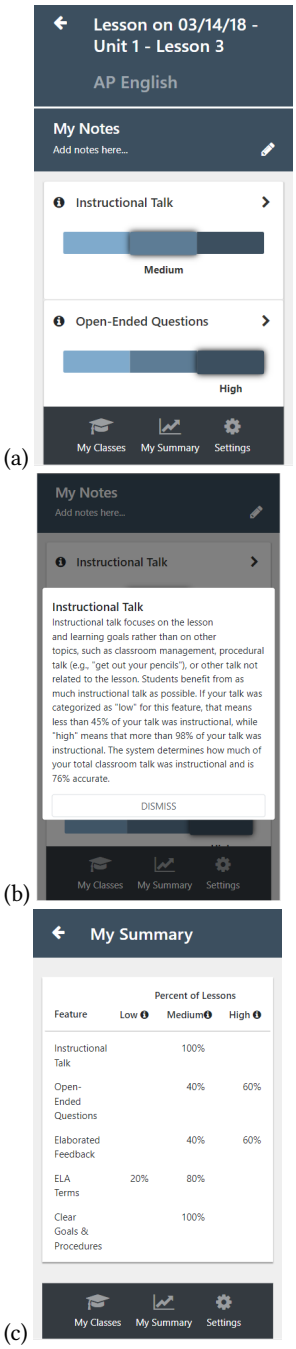


Figure 3: Screenshots of preliminary teacher feedback (a) using session-level ordinal feedback, (b) specific information about the discourse feedback, and (c) a summary of all lessons

4.3 Implications and Applications

Given the diverse and complex needs of individual teachers, there is unlikely to be a one-size-fits-all approach to automated discourse feedback. Each situation will vary in their desired outcomes and ability to collect data. The type and amount of data available will dictate which machine learning methods can be used for automated analysis. The results of the current study suggest that the traditional n-gram and RF approach might be a better choice when training data is limited, but BERT is clearly preferred when training data is abundant. Further given the rapid pace of advances in NLP and deep learning, it is prudent to replicate these analyses with newer models such as the Generative Pre-trained Transformer 3 (GPT-3, [15]) model, which is achieving state-of-the-art results in many NLP tasks.

Beyond modeling, future research is needed to understand the most effective ways to provide teachers with feedback. For example, the level of feedback should take into account the desired insights teachers need in order to improve their practice. Pre-service teachers, for instance, may find value in more detailed utterance-level feedback while experienced teachers may use session-level ordinal feedback to periodically review their classroom discourse. We also need to investigate the potential impacts feedback systems may have on teacher learning. Although there is some initial evidence that teacher feedback can be used to improve student learning outcomes [50], there is a dearth of studies that examine the longitudinal effects of presenting these analytics to teachers. As discussed in [77], it is important to move beyond modeling to better understand how teachers are using the given information to make decisions about their instructional practices and which approaches are effective.

Towards this end, our future work will study the impacts of an automated feedback system proposed in [41] which provides teachers with session-level ordinal feedback. Our initial designs are illustrated in Figure 3, where we opted to provide teachers with session-level ordinal feedback using the BERT models and 15:85 split (mean accuracy of about 70%), along with explanations that clearly communicate model accuracy in the interest of transparency, and a summary of the measures across class sessions (e.g., percent of lessons classified as low for Instructional Talk). After evaluating the feedback designs in user studies, we will investigate whether and how teachers alter their behaviors based on the feedback and identify the best way to pare feedback with other forms of coaching or instructional support. There is also the foundational question of whether this form of data-driven professional development can lead to improvements in teacher discourse and whether this results in improved student achievement, which will entail further development and evaluation.

ACKNOWLEDGMENTS

This research was supported by the National Science Foundation (NSF IIS 1735785). Any opinions, findings, conclusions, or recommendations expressed in this paper are those of the author and do not represent the views of the funding agencies. We thank Sarah Capello, Meghan Dale, Sean Kelly, Amanda Godley, and Patrick Donnelly for their contributions to the data collection and coding.

REFERENCES

[1] Ahn, J. *et al.* 2019. Designing in context: reaching beyond usability in learning analytics dashboard design. *Journal of Learning Analytics.* 6, 2 (Jul. 2019). DOI:https://doi.org/10.18608/jla.2019.62.5.

[2] Alibali, M.W. *et al.* 2014. How teachers link ideas in mathematics instruction using speech and gesture: a corpus analysis. *Cognition and Instruction.* 32, 1 (2014), 65–100. DOI:https://doi.org/10.1080/07370008.2013.858161.

[3] An, P. *et al.* 2020. The ta framework: designing real-time teaching augmentation for k-12 classrooms. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, Apr. 2020), 1–17. DOI:https://doi.org/10.1145/3313831.3376277.

[4] Archer, J. *et al.* 2016. Better Feedback for Better Teaching: A Practical Guide to Improving Classroom Observations.

[5] Aslan, S. *et al.* 2019. Investigating the impact of a real-time, multimodal student engagement analytics technology in authentic classrooms. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19* (New York, New York, USA, 2019), 1–12. DOI:https://doi.org/10.1145/3290605.3300534.

[6] Azevedo, R. and Bernard, R.M. 1995. A meta-analysis of the effects of feedback in computer-based instruction. *Journal of Educational Computing Research.* 13, 2 (1995), 111–127. DOI:https://doi.org/10.2190/9lmd-3u28-3a0g-ftqt.

[7] Ball, D.L. and Cohen, D.K. 1999. Developing practice, developing practitioners: toward a practice-based theory of professional education. *Teaching as the learning profession: Handbook of policy and practice.* G. Sykes and L. Darling-Hammond, eds. Jossey-Bass Inc. 3–32.

[8] Barmaki, R. and Hughes, C.E. 2018. Embodiment analytics of practicing teachers in a virtual immersive environment. *Journal of Computer Assisted Learning.* 34, 4 (Aug. 2018), 387–396. DOI:https://doi.org/10.1111/jcal.12268.

[9] Benjamini, Y. and Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society.* 57, 1 (1995), 289–300.

[10] Blanchard, N. *et al.* 2016. Automatic detection of teacher questions from audio in live classrooms. *Proceedings of the 9th International Conference on Educational Data Mining (EDM 2016)* (2016).

[11] Borko, H. 2004. Professional development and teacher learning: mapping the terrain. *Educational Researcher.* 33, 8 (2004), 3–15. DOI:https://doi.org/10.3102/0013189X033008003.

[12] Bosch, N. *et al.* 2018. Quantifying classroom instructor dynamics with computer vision. *Artificial Intelligence in Education* (2018), 30–42. DOI:https://doi.org/10.1007/978-3-319-93843-1_3.

[13] Bransford, J.D. *et al.* eds. 2000. *How people learn: brain, mind, experience, and school.* National Academy Press.

[14] Brasoveanu, A.M.P. and Andonie, R. 2020. Visualizing transformers for nlp: a brief survey. *24th International Conference Information Visualisation* (2020), 257–266. DOI:https://doi.org/10.1109/IV51561.2020.00051.

[15] Brown, T.B. *et al.* 2020. Language models are few-shot learners. (May 2020).

[16] Cantrell, S. and Kane, T.J. 2013. Ensuring Fair and Reliable Measures of Effective Teaching: Culminating Findings from the MET Project's Three-Year Study.

[17] Chua, Y.H.V. *et al.* 2019. Technologies for automated analysis of co-located, real-life, physical learning spaces. *Proceedings of the 9th International Conference on Learning Analytics & Knowledge* (New York, NY, USA, Mar. 2019), 11–20. DOI:https://doi.org/10.1145/3303772.3303811.

[18] Church, K.W. and Hanks, P. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics.* 16, 1 (1990), 76–83. DOI:https://doi.org/10.3115/981623.981633.

[19] Cook, C. *et al.* 2018. An open vocabulary approach for estimating teacher use of authentic questions in classroom discourse. *Proceedings of the 11th International Conference on Educational Data Mining* (2018).

[20] D'Mello, S.K. *et al.* 2010. Expert tutors' feedback is immediate, direct, and discriminating. *Proceedings of the 23rd International Florida Artificial Intelligence Research Society Conference, FLAIRS-23* (Menlo Park, CA, 2010), 504–509.

[21] D'Mello, S.K. *et al.* 2015. Multimodal capture of teacher-student interactions for automated dialogic analysis in live classrooms. *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction - ICMI '15* (New York, New York, USA, 2015), 557–566. DOI:https://doi.org/10.1145/2818346.2830602.

[22] Devlin, J. *et al.* 2019. BERT: pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (Minneapolis, MN, 2019), 4171–4186. DOI:https://doi.org/10.18653/v1/N19-1423.

[23] Donnelly, P.J. *et al.* 2016. Automatic teacher modeling from live classroom audio. *Proceedings of the 24th Conference on User Modeling, Adaptation, and Personalization (UMAP 2016)* (2016).

[24] Duckworth, A.L. *et al.* 2010. Deliberate practice spells success: why grittier competitors triumph at the national spelling bee. *Social Psychological and Personality Science.* 2, 2 (2010), 174–181. DOI:https://doi.org/10.1177/1948550610385872.

[25] Duke, N. *et al.* 2012. *Reading and writing genre with purpose in a k-8 classroom.* Heinemann.

[26] Ericsson, K.A. 2006. The influence of experience and deliberate practice on the development of superior expert performance. *The Cambridge Handbook of Expertise and Expert Performance.* K.A. Ericsson *et al.*, eds. Cambridge University Press. 685–706.

[27] Ericsson, K.A. *et al.* 1993. The role of deliberate practice in the acquisition of expert performance. *Psychological Review.* 100, 3 (1993), 363–406. DOI:https://doi.org/10.1037/0033-295x.100.3.363.

[28] Fadde, P.J. and Klein, G.A. 2010. Deliberate performance: accelerating expertise in natural settings. *Performance Improvement.* 49, 9 (2010), 5–14. DOI:https://doi.org/10.1002/pfi.

[29] Gamoran, A. and Nystrand, M. 1992. Taking students seriously. *Student Engagement and Achievement in American Secondary Schools.* F.M. Newmann, ed. Teachers College Press. 40–61.

[30] Garet, M.S. *et al.* 2011. Middle School Mathematics Professional Development Impact Study: Findings After the Second Year of Implementation.

[31] Garet, M.S. *et al.* 2008. The Impact of Two Professional Development Interventions on Early Reading Instruction and Achievement.

[32] Goldman, R. *et al.* eds. *Video research in the learning sciences.* Erlbaum.

[33] Graesser, A.C. *et al.* 2009. What is a good question? *Threads of coherence in research on the development of reading ability.* M.G. McKeown and L. Kucan, eds. Guilford Press. 112–141.

[34] Grossman, P. *et al.* 2013. Measure for measure: the relationship between measures of instructional practice in middle school english language arts and teachers' value-added scores. *American Journal of Education.* 119, 3 (2013), 445–470. DOI:https://doi.org/10.1086/669901.

[35] Gwet, K.L. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology.* 61, (2008), 29–48.

[36] Hill, H.C. *et al.* 2008. Mathematical knowledge for teaching and the mathematical quality of instruction: an exploratory study. *Cognition and Instruction.* 26, 4 (2008), 430–511.

[37] Hittner, J.B. *et al.* 2003. A monte carlo evaluation of tests for comparing dependent correlations. *The Journal of General Psychology.* 130, 2 (Apr. 2003), 149–168. DOI:https://doi.org/10.1080/00221300309601282.

[38] Holstein, K. *et al.* 2018. Student learning benefits of a mixed-reality teacher awareness tool in ai-enhanced classrooms. *Artificial Intelligence in Education* (2018), 154–168. DOI:https://doi.org/10.1007/978-3-319-93843-1_12.

[39] Hooreman, R.W. *et al.* 2008. Effects of synchronous coaching in teacher training. *International Journal of Continuing Engineering Education and Life-Long Learning.* 18, 3 (2008), 338. DOI:https://doi.org/10.1504/IJCEELL.2008.018836.

[40] Van Hulse, J. *et al.* 2007. Experimental perspectives on learning from imbalanced data. *Proceedings of the 24th international conference on Machine learning - ICML '07* (New York, New York, USA, 2007), 935–942. DOI:https://doi.org/10.1145/1273496.1273614.

[41] Jensen, E. *et al.* 2020. Toward automated feedback on teacher discourse to enhance teacher learning. *2020 CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2020)* (2020). DOI:https://doi.org/10.1145/3313831.3376418.

[42] Juzwik, M.M. *et al.* 2013. *Inspiring dialogue: talking to learn in the english classroom.* Teachers College Press.

[43] Kelly, S. *et al.* 2018. Automatically measuring question authenticity in real-world classrooms. *Educational Researcher.* 47, 7 (2018), 451–464.

[44] Kelly, S. 2007. Classroom discourse and the distribution of student engagement. *Social Psychology of Education.* 10, 3 (2007), 331–352.

[45] Lemaître, G. *et al.* 2017. Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. *The Journal of Machine Learning Research.* 18, (2017), 559–563. DOI:https://doi.org/{http://www.jmlr.org/papers/volume18/16-365/16-365.pdf}.

[46] Li, Y. and Yang, T. 2018. Word embedding for understanding natural language: a survey. *Guide to Big Data Applications. Studies in Big Data.* S. Srinivasan, ed. 83–104.

[47] Lugrin, J.-L. *et al.* 2016. Breaking bad behaviors: a new tool for learning classroom management using virtual reality. *Frontiers in ICT.* 3, (Nov. 2016). DOI:https://doi.org/10.3389/fict.2016.00026.

[48] Lugrin, J.-L. *et al.* 2018. VR-assisted vs video-assisted teacher training. *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)* (Mar. 2018), 625–626. DOI:https://doi.org/10.1109/VR.2018.8446312.

[49] McKenney, S. and Mor, Y. 2015. Supporting teachers in data-informed educational design. *British Journal of Educational Technology.* 46, 2 (Mar. 2015), 265–279. DOI:https://doi.org/10.1111/bjet.12262.

[50] Molenaar, I. and Knoop-van Campen, C.A.N. 2019. How teachers make dashboard information actionable. *IEEE Transactions on Learning Technologies.* 12, 3 (Jul. 2019), 347–355. DOI:https://doi.org/10.1109/TLT.2018.2851585.

[51] Murphy, P.K. *et al.* 2009. Examining the effects of classroom discussion on students' comprehension of text: a meta-analysis. *Journal of Educational Psychology.* 101, 3 (2009), 740–764. DOI:https://doi.org/10.1037/a0015576.

[52] Ndukwe, I.G. and Daniel, B.K. 2020. Teaching analytics, value and tools for teacher data literacy: a systematic and tripartite approach. *International Journal of Educational Technology in Higher Education.* 17, 1 (Dec. 2020), 22. DOI:https://doi.org/10.1186/s41239-020-00201-6.

[53] Nystrand, M. *et al.* 1997. Opening dialogue: understanding the dynamics of language and learning in the english classroom. Teachers College Press.

[54] Olney, A.M. *et al.* 2017. Assessing the dialogic properties of classroom discourse: proportion models for imbalanced classes. *Proceedings of the 10th International Conference on Educational Data Mining* (2017), 162–167.

[55] Pan, S.J. and Yang, Q. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering.* 22, 10 (Oct. 2010), 1345–1359. DOI:https:

//doi.org/10.1109/TKDE.2009.191.

[56] Pedregosa, F. *et al.* 2011. Scikit-learn: machine learning in python. *Journal of Machine Learning Research*. 12, (2011), 2825–2830. DOI:https://doi.org/10.1007/s13398-014-0173-7.2.

[57] Prieto, L.P. *et al.* 2016. Teaching analytics: towards automatic extraction of orchestration graphs using wearable sensors. *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge - LAK '16* (New York, New York, USA, 2016), 148–157. DOI:https://doi.org/10.1145/2883851.2883927.

[58] Raudenbush, S.W. *et al.* 1993. Higher order instructional goals in secondary schools: class, teacher, and school influences. *American Educational Research Journal*. 30, 3 (Jan. 1993), 523–553. DOI:https://doi.org/10.3102/00028312030003523.

[59] Resnick, L.B. and Schantz, F. 2015. Re-thinking intelligence: schools that build the mind. *European Journal of Education*. 50, 3 (2015), 340–349. DOI:https://doi.org/10.1111/ejed.12139.

[60] Robin, X. *et al.* 2011. PROC: an open-source package for r and s+ to analyze and compare roc curves. *BMC Bioinformatics*. 12, 1 (Dec. 2011), 77. DOI:https://doi.org/10.1186/1471-2105-12-77.

[61] Saon, G. *et al.* 2015. The ibm 2015 english conversational telephone speech recognition system. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* (Dresden, Germany, 2015), 3140–3144. DOI:https://doi.org/10.21437/Interspeech.2016-1460.

[62] Scheeler, M.C. *et al.* 2012. Effects of immediate feedback delivered via webcam and bug-in-ear technology on preservice teacher performance. *Teacher Education and Special Education: The Journal of the Teacher Education Division of the Council for Exceptional Children*. 35, 1 (Feb. 2012), 77–90. DOI:https://doi.org/10.1177/0888406411401919.

[63] Schuster, M. and Nakajima, K. 2012. Japanese and korean voice search. *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Mar. 2012), 5149–5152. DOI:https://doi.org/10.1109/ICASSP.2012.6289079.

[64] Schwendimann, B.A. *et al.* 2017. Perceiving learning at a glance: a systematic literature review of learning dashboard research. *IEEE Transactions on Learning Technologies*. 10, 1 (Jan. 2017), 30–41. DOI:https://doi.org/10.1109/TLT.2016.2599522.

[65] Sergis, S. and Sampson, D.G. 2017. Teaching and learning analytics to support teacher inquiry: a systematic literature review. *Learning Analytics: Fundaments, Applications, and Trends. Studies in Systems, Decision and Control*. 25–63.

[66] Shernoff, D.J. *et al.* 2016. Student engagement as a function of environmental complexity in high school classrooms. *Learning and Instruction*. 43, (2016), 52–60. DOI:https://doi.org/10.1016/j.learninstruc.2015.12.003.

[67] Shernoff, D.J. *et al.* 2003. Student engagement in high school classrooms from the perspective of flow theory. *School Psychology Quarterly*. 18, 2 (2003), 158.

[68] Shute, V.J. 2008. Focus on formative feedback. *Review of Educational Research*. 78, 1 (2008), 153–189. DOI:https://doi.org/10.3102/0034654307313795.

[69] Stone, C. *et al.* 2019. Utterance-level modeling of indicators of engaging classroom discourse. *The 12th International Conference on Educational Data Mining* (Montreal, Canada, 2019), 420–425.

[70] Stone, C. 2019. Utterance-level modeling of indicators of engaging classroom discourse (draft). *Educational Data Mining* (2019).

[71] Suresh, A. *et al.* 2019. Automating analysis and feedback to improve mathematics' teachers' classroom discourse. *Proceedings of the Ninth Symposium on Educational Advances in Artificial Intelligence (EAAI)* (2019).

[72] Suresh, A. *et al.* 2018. Using deep learning to automatically detect talk moves in teachers'mathematics lessons. *Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018*. (2018), 5445–5447. DOI:https://doi.org/10.1109/BigData.2018.8621901.

[73] Taylor, B.M. *et al.* 2003. The influence of teacher practices that encourage cognitive engagement in literacy learning. *The Elementary School Journal*. 104, 1 (2003), 3–28.

[74] TNTP 2015. The Mirage: Confronting the Hard Truth About Our Quest for Teacher Development.

[75] Vaswani, A. *et al.* 2017. Attention is all you need. *Advances in neural information processing systems* (2017), 5998–6008.

[76] Verbert, K. *et al.* 2013. Learning analytics dashboard applications. *American Behavioral Scientist*. 57, 10 (2013), 1500–1509. DOI:https://doi.org/10.1177/0002764213479363.

[77] Verbert, K. *et al.* 2014. Learning dashboards: an overview and future research opportunities. *Personal and Ubiquitous Computing*. 18, 6 (2014), 1499–1514. DOI:https://doi.org/10.1007/s00779-013-0751-2.

[78] Wang, Z. *et al.* 2013. Using the lena in teacher training: promoting student involvement through automated feedback. *Unterrichtswissenshaft*. 4, (2013), 290–305.

[79] Wolf, T. *et al.* 2019. HuggingFace's transformers: state-of-the-art natural language processing. (Oct. 2019).