frontiers | Frontiers in Education

TYPE Methods
PUBLISHED 15 September 2023
DOI 10.3389/feduc.2023.1250846

OPEN ACCESS

EDITED BY
Julius Nganji,
University of Toronto, Canada

REVIEWED BY
Bob Edmison,
Virginia Tech, United States
Rosanna Yuen-Yan Chan,
The Chinese University of Hong Kong, China

*CORRESPONDENCE
Michael Yee
✉ myee@ll.mit.edu

RECEIVED 30 June 2023
ACCEPTED 29 August 2023
PUBLISHED 15 September 2023

CITATION
Yee M, Roy A, Perdue M, Cuevas C, Quigley K, Bell A, Rungta A and Miyagawa S (2023) AI-assisted analysis of content, structure, and sentiment in MOOC discussion forums. *Front. Educ.* 8:1250846. doi: 10.3389/feduc.2023.1250846

AI-assisted analysis of content, structure, and sentiment in MOOC discussion forums

Michael Yee[1]*, Anindya Roy[2], Meghan Perdue[2], Consuelo Cuevas[1], Keegan Quigley[1], Ana Bell[3], Ahaan Rungta[2] and Shigeru Miyagawa[4]

[1]Artificial Intelligence Technology Group, MIT Lincoln Laboratory, Lexington, MA, United States, [2]Open Learning, Massachusetts Institute of Technology, Cambridge, MA, United States, [3]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, United States, [4]Department of Linguistics and Philosophy, Massachusetts Institute of Technology, Cambridge, MA, United States

Discussion forums are a key component of online learning platforms, allowing learners to ask for help, provide help to others, and connect with others in the learning community. Analyzing patterns of forum usage and their association with course outcomes can provide valuable insight into how learners actually use discussion forums, and suggest strategies for shaping forum dynamics to improve learner experiences and outcomes. However, the fine-grained coding of forum posts required for this kind of analysis is a manually intensive process that can be challenging for large datasets, e.g., those that result from popular MOOCs. To address this issue, we propose an AI-assisted labeling process that uses advanced natural language processing techniques to train machine learning models capable of labeling a large dataset while minimizing human annotation effort. We fine-tune pretrained transformer-based deep learning models on category, structure, and emotion classification tasks. The transformer-based models outperform a more traditional baseline that uses support vector machines and a bag-of-words input representation. The transformer-based models also perform better when we augment the input features for an individual post with additional context from the post's thread (e.g., the thread title). We validate model quality through a combination of internal performance metrics, human auditing, and common-sense checks. For our Python MOOC dataset, we find that annotating approximately 1% of the forum posts achieves performance levels that are reliable for downstream analysis. Using labels from the validated AI models, we investigate the association of learner and course attributes with thread resolution and various forms of forum participation. We find significant differences in how learners of different age groups, gender, and course outcome status ask for help, provide help, and make posts with emotional (positive or negative) sentiment.

KEYWORDS

MOOCs, discussion forums, forum posts, natural language processing, text classification, machine learning, transformers, artificial intelligence

1. Introduction

Massive Open Online Courses (MOOCs) are tremendous educational resources for learners seeking to educate themselves or gain new skills. A key component of MOOCs is the discussion forums, a place for learners to engage in conversation, provide and receive help, and establish a learning community. Although studies have not found evidence that

strong social networks are formed in MOOC discussion forums (Gillani and Eynon, 2014; Boroujeni et al., 2017; Wise et al., 2017), they do create learning communities where learners can get support (Poquet and Dawson, 2015). Additionally, studies have shown that giving and receiving help from student peers increases learning gains (Topping, 2005; Yamarik, 2007). Others have shown that engagement in discussion forums in MOOCs is linked to learner retention in the course (Houston et al., 2017; Poquet et al., 2018). However, analyzing discussion forum use is challenging due to the vast amount of unstructured data and the complexity of the interactions. Artificial Intelligence (AI) offers an opportunity to overcome these difficulties and provide valuable insights into the learning process within discussion forums. This research will address the following questions:

- RQ1: Can we use AI algorithms to tag forum posts along category, structure, and emotion dimensions as reliably as human coders? If the answer to RQ1 is yes, we can use AI-generated tags to answer the following research questions:
- RQ2: How are the tags distributed in the forums and within threads? How are the tags related to each other?
- RQ3: How are learner attributes and course attributes associated with the forum participants? Specifically, how are these attributes associated with whether a thread started by a learner got resolved, and what is the likelihood of such a learner posting a comment of a certain type?

2. Related work

2.1. Online education and forums

Researchers have sought to identify patterns in discussion forum usage that could shed light on which learners are using the forums, and how they are using them. Many studies have shown that there are variations of forum usage, with some learners using the forums more than others, though the most consistent group is active learners interested in completing the course (Huang et al., 2014; Almatrafi and Johri, 2018; Moreno-Marcos et al., 2018). Research into forum usage over time has found that the number of learners using the forum and overall quantity of posts diminish through the course run (Brinton et al., 2014). However, other studies (Wong et al., 2015; Galikyan et al., 2021) found that learners used progressively higher cognitive levels as they advanced throughout the course.

Forums support a broad range of social activities in the course, such as small talk, questions about the logistics, help-seeking and help-giving behaviors, and content-based discussion. Studies have investigated the impact of participating in content vs. non-content-based discussion threads, and found that engagement in the forums was positively correlated with course performance, regardless of post type (Wise et al., 2017; Wise and Cui, 2018). Boroujeni et al. (2017) found that the forums play a particularly useful role in content-triggering discussions, especially for help-seekers. Some research has been conducted to seek to unpack the degree to which the forums are effective for learners who need additional support with the content (Kim and Kang, 2014). Yang et al. (2015) looked specifically at the impact of unresolved confusion or help-seeking on learners. They found that expressing confusion in the forums was negatively correlated with course retention, though this was mediated by receiving support and resolving the confusion.

Relatedly, there has been interest in whether the emotional sentiment of a learner's post could predict their retention in the course. Ezen-Can et al. (2015) sought to understand the relationship between the emotional sentiment of a learner's posts and their persistence in the course by creating a sentiment score for each learner based on all their posts. They looked at three distinct MOOCs and found that the patterns differed for each course, but for the computer science course, a significant emotional response (positive or negative) was correlated with a higher dropout rate. Wen et al. (2014) sought to model learner dropout rates based on a sentiment analysis of learner discussion forums posts and usage type and found that positive and actively engaged learners were most likely to complete the course, followed by active and negative.

MOOC forum posts have also been analyzed in connection to participants' demographic attributes, such as gender and age. In a study conducted by Swinnerton et al. (2017), it was discovered that older learners are more likely to post comments. Another study by Huang et al. (2014) compared forum superposters to ordinary participants in terms of age and gender. The findings revealed that superposters tend to be older than the average forum users, and there is a small but statistically significant over-representation of women among superposters, while there are generally more male forum participants. Gender differences in MOOC forum posts have been explored by John and Meinel (2020) by analyzing the types of questions asked by men and women, as well as the categories of responses these questions generated. Although they did not find statistical significance between the question types and gender, they observed that male learners tended to participate more in longer discussions.

2.2. Automated forum post classification

Studying forum data requires characterizing individual posts along one or more dimensions, such as topic, activity, degree of confusion, and sentiment. This data enrichment process can require a significant investment of time and/or money especially if the dataset is large. However, each dimension can be formulated as a multi-class text classification problem, and natural language processing (NLP) techniques can be used to help automate or semi-automate the labeling process.

Early applications of machine learning to text classification involved representing textual content as a bag of words (or longer n-grams), with terms weighted by their term frequency–inverse document frequency (TF-IDF) weights, and training shallow machine learning models such as support vector machine (SVM), logistic regression, and random forest classifiers (Schütze et al., 2008).

Like other domains, NLP was revolutionized by deep learning (LeCun et al., 2015). Pretrained word embeddings

such as word2vec (Mikolov et al., 2013a,b) and GloVe (Pennington et al., 2014) introduced a powerful alternative to the bag-of-words representation, and a variety of neural network architectures such as multilayer perceptrons (MLPs), recurrent neural networks (RNNs), convolutional neural networks (CNNs), and transformer models have achieved state of the art performance on many NLP tasks [see Minaee et al. (2021) and Li et al. (2022) for overviews of text classification].

The same progression from traditional (shallow) machine learning methods to deep learning has taken place in the domain of forum post classification (Ahmad et al., 2022). Since the present work explores three distinct classification tasks (category, structure, and emotion), the use of transformer-based models, and strategies for reducing human annotation burden, we highlight relevant work in these areas.

The classification task studied in Ntourmas et al. (2019, 2021) is nearly identical to our category task; Ntourmas et al. (2019) classified starting posts as content-related, logistics-related, or other using an SVM with bag-of-words representation, while Ntourmas et al. (2021) applied a decision tree to a TF-IDF weighted bag-of-words representation along with additional features derived from a seeded topic modeling technique. Sentiment analysis is a common forum post classification task studied in, e.g., Bakharia (2016), Chen et al. (2019); Clavié and Gal (2019), Li et al. (2019), and Capuano et al. (2021) . Sentiment is also one of the six dimensions (question, opinion, answer, sentiment, urgency, and confusion) included in the Stanford MOOCPosts dataset (Agrawal et al., 2015), slices of which have been used by many works, e.g., Bakharia (2016), Chen et al. (2019), Clavié and Gal (2019), Guo et al. (2019), Sun et al. (2019), and Alrajhi et al. (2020). For automatic analysis of thread structure, Sun et al. (2016) classified posts within a thread according to dialogue acts (question, answer, resolution, reproduction, other) and also whether one post contains an immediate follow-up discussion of another using conditional random fields (CRFs). Joksimović et al. (2019) used an unsupervised approach combining hidden Markov models (HMMs) and Latent Dirichlet Allocation (LDA) to discover and analyze speech act categories. Fisher et al. (2015) trained an HMM-like latent variable model using weak supervision to classify whether one post is a direct response to another.

Recent works spanning numerous forum post classification tasks have incorporated pretrained transformers such as BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018), e.g., Clavié and Gal (2019), Li et al. (2019), Sha et al. (2021), Zou et al. (2021), and Lee et al. (2022). BERT is a promising technique since it is often possible to fine-tune models pretrained on large unlabeled datasets using a limited amount of labeled task-specific data.

Other works addressing the problem of annotation burden include Ntourmas et al. (2021), which investigated how many weeks of labeled data are required to perform reliably when applied to the remaining weeks of the MOOC and finds that maximum performance is achieved by week 3, and Chen et al. (2019), which developed a semi-supervised learning approach based on co-training and found that it outperforms traditional and deep learning baselines while starting with only 30% of the dataset labeled.

3. Methodology

Our overall approach is to create a training set by manually annotating a relatively small subset of the full forum post dataset, train AI models on the labeled data, apply the best-performing models to the unlabeled remainder of the data, and finally use the fully labeled dataset (containing AI-generated predictions plus a small fraction of human-supplied labels) to answer downstream education-related research questions.

To answer RQ1, we implement the following AI-assisted labeling process:

- **Step 1**: Develop initial coding guidelines
- **Step 2**: Annotate new subset of unlabeled posts
- **Step 3**: Measure inter-annotator agreement and adjudicate labels when annotators disagree
- **Step 4**: Refine coding guidelines and return to Step 2 (if necessary due to low agreement)
- **Step 5**: Train AI models on annotated data
- **Step 6**: Evaluate model performance with internal metrics and comparison to human performance, and return to Step 2 to increase training set size (if necessary)
- **Step 7**: Apply best models to unlabeled data
- **Step 8**: Audit subset of model predictions and return to Step 2 to increase training set size (if necessary)

Assuming the model predictions pass the final human audit, the fully labeled dataset generated in Step 7 is then used to answer RQ2 and RQ3. We use exploratory data analysis techniques at both the post and thread level to answer RQ2: using Jaccard similarity to measure co-occurrence of tags from the three dimensions (category, structure, and emotion), and aggregating posts within threads to analyze how tag distributions vary with thread length. For RQ3, we use logistic regression to investigate the association of forum behavior with learner and course attributes.

4. MOOC dataset

For this work we studied 11 course instances of two sequential introductory Python MOOCs (referred to as Python-1 and Python-2): six of these eleven courses were Python-1, and the remaining five were Python-2. The courses ran between Spring 2018 and Summer 2021—which includes a total of five course instances that operated during the COVID-19 pandemic. Since these courses were all offered online, they did not undergo any change in logistics during the pandemic, and their content remained the same over this period. All the courses were 9 weeks long and instructor-paced, with defined start and end dates. Approximately 387,000 learners enrolled for these 11 courses, about 6% of whom paid for full access and the opportunity to earn a certificate ("verified learners"), and ∼53% of the verified learners went on to earn a certificate. The learners along with their instructors and community teaching assistants (TAs) generated ∼82000 posts. Additionally, we have

detailed records of how the learners interacted with various other course materials such as problems and videos, but we do not take that into account for the purpose of this work. For many of these learners we have voluntarily supplied demographic information such as their age, gender, and level of education, and we include that in our work as a representative sample. In addition, we could infer the country where most learners accessed the courses, and note the country's economic category (e.g., high-income country, lower-middle-income country, etc.). In this work, we use learners' verification and certification status, as well as when a particular course ran (i.e., before or during the pandemic) where relevant. For more information on the demographic details of the learners, we refer our readers to the authors' previous work in Roy et al. (2022) and Yee et al. (2022).

5. Annotation process

To create the training dataset for AI model development, our team members ultimately labeled a collection of 950 forum posts randomly sampled from the 11 Python MOOCs, with approximately an equal share coming from each course. Each team member labeled an initial set of 50 posts. The initial labels selected in the project were based on previous research conducted in this area analyzing forum posts (Brinton et al., 2014; Boroujeni et al., 2017; Wise et al., 2017; Galikyan et al., 2021) and optimized for the research questions. The labels were finalized by reviewing sample codes with all coders present and discussing how the labels would apply to the data. Once the group came to a consensus on the application, the labels and definitions were drafted into coding guidelines that were distributed to all coders. The labels for each task were:

- **Category**: Logistics, Content, Emotional/Commentary
- **Structure**: Question, Suggestion/Explanation, Follow Up/Follow Up Question, Resolution, Comment/Response
- **Emotion**: Positive, Negative, Neutral

Sentiment analysis has been studied with various formulations ranging from binary classification (positive/negative) to multi-point scales (Pang and Lee, 2008; Zhang et al., 2018). We use three labels (positive/negative/neutral) to support our downstream analysis.

After labeling, we held an adjudication session where we discussed disagreements, decided on adjudicated (consensus) answers, and refined the coding guidelines. When two or more coders disagreed on a label, a group of three or more coders reviewed the data together and came to a consensus on the correct code to use, and any changes in the interpretation of labeling was detailed in the coding guidelines. The team labeled three additional sets of posts (of size 250, 300, and 350) with two team members annotating each post, followed by additional adjudication sessions. The additional rounds were required to improve either coding understanding and consistency or current model performance (Steps 4 and 6). The final coding guidance is given in Table 1.

We computed Krippendorf's alpha (Krippendorff, 1970) to assess consistency across annotators (inter-annotator agreement) and the quality of the coding guidelines. For the Category, Structure, and Emotion tasks, we achieved alphas of 0.632, 0.634, and 0.532, respectively. These values are lower than what is generally considered to be very good agreement (alpha $\geq$ 0.8). However, these final values span multiple labeling rounds and evolving annotation team membership. Since we adjudicated all disagreements among annotators for the entire training set, we considered the final labels to be high quality and sufficient for model training and downstream analysis.

In Figure 1, we present the relative percentages of the classes within the three tasks as coded by the human annotators, as well as those predicted by the AI-assisted labeling process (only the subset of the posts made by the learners alone is presented here). The full predicted dataset, which includes posts from instructors and TAs, is closer to the distributions of the manually annotated set. Note that the distributions of the classes within each task revealed moderate to severe class imbalance, e.g., negative posts only made up 5% of the final annotated dataset. This was a key challenge in developing accurate models.

6. Model design and performance

In this section, we describe data preprocessing approaches, model architectures, and training hyperparameters that we explored to develop models capable of accurately labeling the Category, Structure, and Emotion dimensions for unlabeled forum posts.

6.1. Data preprocessing

To address the class imbalance and limited size of the labeled dataset, multiple dataset preprocessing steps were tested. First, contextual information about each forum post was added as input to the model. Discussions in forums on the edX platform take place in threads with the following structure: thread title, initial post, zero or more response posts, and zero or more comments on each response post. For each forum post in the dataset, we reconstructed the containing thread (using forum database tables persisted by edX) and extracted the following thread context features:

- `thread_title`
- `post_type`: initial_post, response_post, or comment
- `num_responses`: number of responses in thread
- `response_position`: position in response list for the response containing this post (if post is response or comment)
- `num_response_comments`: number of comments on the response containing this post (if post is response or comment)
- `comment_position`: position in comment list for the response containing this comment (if post is comment)
- `original_poster`: whether post is authored by the same author as thread's initial post

To aid our annotation process, we condensed these features into a short textual form. For example, a post with context "Response 1/2; Num Comments 7" means it was the first response out of two and it received a total of seven comments.

TABLE 1 Coding guidance for the Category, Structure, and Emotion forum post classification tasks.

Task	Code	Definition
Category	Logistics	Post relates to logistics of using the platform/accessing materials in the course. Typically noting that a problem set isn't working, there is a typo in the materials, their submission didn't go through for some reason, or they are asking for an extension.
Category	Content	Post relates to the content of the course itself. Typically asking a question about the material, clarifying understanding, asking for help with a problem, requesting additional materials to understand the concept, remarking on the videos/materials, etc.
Category	Emotional/Commentary	Post conveys the learners feelings about the class, problem, or experience. Often can be commenting on the pace/difficulty of the course, encouraging others in the class to keep going, etc.
Structure	Question	Post is a question—specifically seeking help for an issue in the course. Often this could be asking for help due to a logistics issue, or help with understanding the content and completing the assignments.
Structure	Suggestion/Explanation	Post is a comment to another learners question post, giving suggestions for how to solve their issue, or trying to explain what the OP doesn't understand.
Structure	Follow Up/Follow Up Question	Post is a follow up to a suggestion/explanation post, but not a resolution. Typically learners would make a follow up post if they attempted the suggestion/explanation and it didn't work, or they want to clarify something.
Structure	Resolution	Post indicates that the question has been resolved. Typically noting that either the suggestion/explanation was sufficient and successful, or noting that it wasn't successful but the learner has given up on the problem or issue.
Structure	Comment/Response	Post is outside the question/suggestion/ follow up/resolution stream. Often general or specific comments about observations, thoughts, feelings.
Emotion	Positive	Post is overwhelmingly positive in sentiment.
Emotion	Negative	Post is overwhelmingly negative in sentiment.
Emotion	Neutral	Post is overwhelmingly neutral in sentiment.

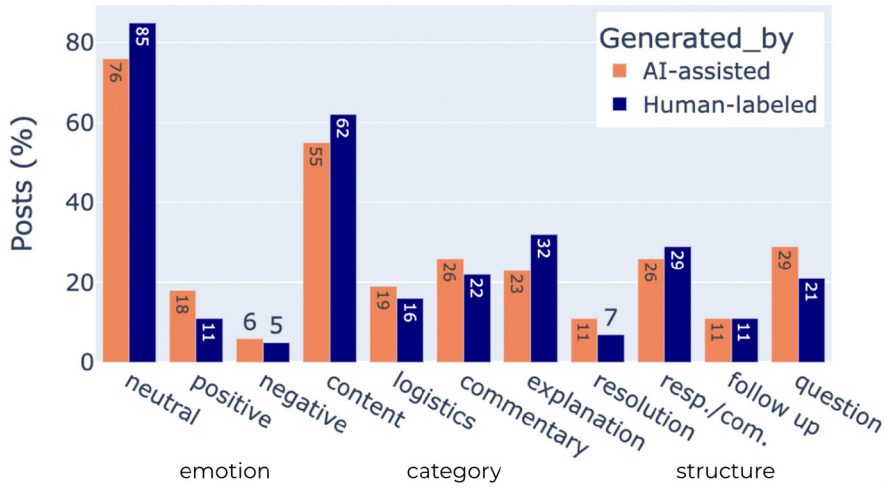

FIGURE 1
Distribution (%) of different classes under each task labeled by AI and human annotators. Classes under each task sum up to 100%. The percentages are obtained from 950 posts labeled by human coders, and from >61,000 learner posts labeled by the trained AI model.

We tested adding this information as input to the model in multiple ways: by concatenating the textual context information to the forum post body; and by constructing a short vector of context features, processing these with a multilayer perceptron (MLP), and concatenating the context embedding with the body embedding before classification.

Because of extreme class imbalance, especially in the case of Negative emotional content, we looked for ways to augment our labeled dataset. Two common methods employed in NLP tasks for dataset augmentation are backtranslation and synonym replacement, both of which we tested. In backtranslation, monolingual data is auto-translated to a foreign language and back to the original language, leading to slight variations in wording but preserving sentiment and content (Sennrich et al., 2015). Synonym replacement replaces random words in the post with their synonyms, with some rules guiding which words can or cannot be replaced (Niu and Bansal, 2018).

Because backtranslation augmentation creates just a single alternative text body from a forum post and due to computational constraints, we created three augmented text bodies for each labeled forum post, and could randomly draw from the four bodies during training. These augmented bodies (two backtranslation, one synonym replacement) were created using the NLPAug Python package (Ma, 2019), with backtranslation performed through two intermediary languages, German and Russian, using Facebook FAIR's pretrained WMT19 News Translation models (Ng et al., 2019). During training, we tested *no augmentation* (`none`), random backtranslation (`backtranslation`), and random backtranslation/synonym replacement (`random`). For `backtranslation` and `random` augmentations, the augmentation was applied to the forum post with a probability of 0.5.

Beyond adding augmentations and additional context vectors, we also tested using weighted random sampling to draw samples from the training dataset for each training batch. This effectively evened the number of training examples from each class. Together with augmentations, we hypothesized that weighted random sampling might lead to higher accuracy across all classes, including those that suffered from severe class imbalance.

6.2. Model architecture

We used a deep learning model consisting of a pretrained BERT encoder (Vaswani et al., 2017; Devlin et al., 2018), and a small MLP classification head that operated on the BERT embeddings. Our classification head consisted of a linear layer with hidden dimension size $h = 128$, ReLU activation, dropout regularization with probability 0.1, and another linear layer projecting from $h \rightarrow$ `num_classes`. If a context vector was provided to the model, we processed this input vector with an MLP, consisting of a linear layer with hidden dimension 32, ReLU, dropout with probability 0.1, a second linear layer with hidden dimension 32, ReLU, dropout ($p = 0.1$), and a linear projection to a context embedding vector of size 16. This context embedding vector was then appended to the BERT embedding before classification by the classification head.

We tested two pretrained BERT models from Devlin et al. (2018), `bert-base-cased` (L = 12, H = 768, Total Params = 110M) and `bert-large-cased` (L = 24, H = 1024, Total Params = 340M). We also tested the `BERT-tweet-eval-emotion` model from HuggingFace, pretrained on the `tweet_eval` dataset (Rosenthal et al., 2017; Barbieri et al., 2020; Schmid, 2021), for its performance on our emotion classification task.

For our loss function, we experimented with cross entropy loss (CE) and soft cross entropy loss (SCE) based on estimated class probabilities. For CE, we used the unanimous consensus labels or adjudicated labels (from resolved disagreements) as our targets (see Section 5). For SCE, we estimated class probabilities from the annotation and adjudication process by normalizing the individual annotator labels as "votes." The adjudicated labels were also included as a vote when available. This produced estimates of class likelihood between $[0, 1]$, which were used as soft targets for our model. We trained all models for 10 epochs, and used an AdamW optimizer (LR = 2e−5, ϵ = 1e−8) with linear warmup (2 epochs) and linear decay (8 epochs).

TABLE 2 **Best model hyperparameters (architecture) and performance for the Structure, Category, and Emotion tasks. Mean accuracy and macro F1 scores across the five test sets from cross-validation are reported, along with standard deviations.**

	Structure	Category	Emotion
Augmentation	none	random	backtranslation
Loss function	sce	sce	ce
BERT model	large	large	tweet-eval
Batch size	8	8	16
Context	textual*	vector	none
Accuracy	0.76 ± 0.03	0.82 ± 0.02	0.87 ± 0.03
Macro F1	0.73 ± 0.03	0.77 ± 0.03	0.67 ± 0.08

For Context, textual* indicates textual context format, without a weighted random sampler.

While BERT and other deep learning transformer-based models have dominated the NLP domain in recent years, more traditional ML and NLP tools like SVMs continue to work sufficiently well for certain tasks. We compared our BERT models to SVM "bag-of-words" approaches, establishing a baseline performance level for the three tasks. To construct these models, we used a count vectorizer to transform each text input into a matrix of token counts, transformed this matrix using TF-IDF weighting, and finally fit an SVM to these transformed features and class labels. The SVM used a maximum of 1000 iterations, hinge loss with L2 penalty, and α = 1e−3. We tested the SVM with each of our dataset preprocessing options, including augmentations, weighted random sampling, and appending the context and thread title to the post body before input to the model.

6.3. Model experiments

To assess the performance of each model architecture on the Category, Structure, and Emotion tasks, we followed a five-fold cross-validation procedure in our experiments. In this procedure, we split the data into five equally sized folds, using four-folds for training the model, and one-fold as a "holdout set" for evaluating the model. Since this holdout set has not been used during training of the model, predictions on its elements can be used to estimate the performance of the model on the remaining unlabeled data. For each model architecture, we trained five distinct models during five-fold cross-validation, with each model trained with a different fold held out during training. Using this procedure, we obtained five independent estimates of the performance of the model's architecture. Aggregating these estimates (Nadeau and Bengio, 1999), we obtained the mean performance and variance estimate across the entire labeled dataset, and could compare model architectures. Additionally, cross-validation methods allowed us to use ensemble methods to predict labels during inference.

For the deep learning experiments with BERT, we use early stopping to guard against overfitting to the training set (Prechelt, 2012). After removing a holdout fold from the training data during cross-validation, we perform a further split for deep learning

TABLE 3 Model comparison for the Emotion task.

Model	Augmentation	BERT model	Loss	Thread context	Accuracy	Macro F1
Annotator A	—	—	—	—	0.919	0.823
Annotator B	—	—	—	—	0.780	0.735
SVM	random	—	—	textual	0.820 ± 0.022	0.505 ± 0.057
BERT	**backtranslation**	**BERT-tweet-eval-emotion**	**ce**	**none**	**0.873 ± 0.026**	**0.671 ± 0.081**
BERT	none	BERT-tweet-eval-emotion	ce	none	0.860 ± 0.035	0.613 ± 0.095
BERT	backtranslation	bert-base-cased	ce	none	0.862 ± 0.032	0.588 ± 0.071
BERT	backtranslation	BERT-tweet-eval-emotion	ce	textual	0.876 ± 0.015	0.640 ± 0.067
BERT	backtranslation	BERT-tweet-eval-emotion	sce	none	0.885 ± 0.021	0.653 ± 0.078

TABLE 4 Model comparison for the Structure task.

Model	Augmentation	BERT model	Loss	Thread context	Accuracy	Macro F1
Annotator A	—	—	—	—	0.924	0.905
Annotator B	—	—	—	—	0.780	0.641
SVM	random	—	—	textual	0.596 ± 0.050	0.538 ± 0.049
BERT	**none**	**bert-large-cased**	**sce**	**textual***	**0.758 ± 0.033**	**0.731 ± 0.030**
BERT	random	bert-large-cased	sce	textual*	0.749 ± 0.025	0.711 ± 0.029
BERT	none	bert-base-cased	sce	textual*	0.689 ± 0.041	0.625 ± 0.059
BERT	none	bert-large-cased	sce	textual	0.734 ± 0.023	0.704 ± 0.021
BERT	none	bert-large-cased	ce	textual*	0.738 ± 0.013	0.670 ± 0.052

TABLE 5 Model comparison for the Category task.

Model	Augmentation	BERT model	Loss	Thread context	Accuracy	Macro F1
Annotator A	—	—	—	—	0.939	0.920
Annotator B	—	—	—	—	0.819	0.744
SVM	none	—	—	textual	0.742 ± 0.027	0.679 ± 0.031
BERT	**random**	**bert-large-cased**	**sce**	**vector**	**0.818 ± 0.018**	**0.768 ± 0.026**
BERT	none	bert-large-cased	sce	vector	0.787 ± 0.029	0.730 ± 0.051
BERT	random	bert-base-cased	sce	vector	0.789 ± 0.035	0.735 ± 0.044
BERT	random	bert-large-cased	sce	none	0.783 ± 0.036	0.728 ± 0.042
BERT	random	bert-large-cased	ce	vector	0.779 ± 0.060	0.729 ± 0.055

In Tables 3–5, the model with the best mean Macro F1 score (bolded) was selected as the best model. Annotator Accuracy and Macro F1 is a comparison of individual annotators' labels and the adjudicated labels (for the subset of the dataset that each annotator labeled). The annotators with the best and worst Macro F1 scores are shown. The adjudication process was partially dependent on the individual annotators' labels, so the annotator performance should be considered a *high* target for model performance. For Thread context, textual* indicates textual context format, without a weighted random sampler.

experiments, reserving 10% of the training data as the validation dataset. This validation data is also withheld from training, but after every training epoch, labels for the validation dataset are predicted by the model and the loss is computed. The model from the training epoch with the lowest validation loss is selected as the best model for that fold, and its expected performance is computed using the holdout fold (functioning as the test set).

We conduct a grid search for each of the Category, Structure, and Emotion tasks over various parameters used to construct a model architecture, finding the best-performing model architecture after training and testing using the five-fold cross-validation scheme. For the grid search over deep learning models using BERT, we sweep over augmentation (`none, backtranslation, random`), loss function (`cross entropy, soft cross entropy`), pretrained BERT model (`bert-base-cased/base, bert-large-cased/large,` as well as `BERT-tweet-eval-emotion/tweet` for the Emotion

task only), and batch size (`8, 16`). For each of the gridded parameters, we also experiment with four different configurations of contextual information inputs: information encoded into a six-element context vector, processed with a small MLP, and concatenated to BERT embedding (`vector`); contextual information concatenated with forum post in a structured way before input to the BERT model (`textual`); `textual` without a weighted random sampler to even the number of examples from each class during training (`textual*`); and no contextual information added (`none`). This search produced a series of 144 Emotion and 96 Structure/Category model architectures to compare (see Section 6.4).

We also conduct a grid search for the SVM-based models, sweeping over thread context (`textual`, as above, or `none`), weighted class sampling (`weighted, unweighted`), and text augmentation (`none, backtranslation, random`), producing 12 candidate model architectures for each classification task.

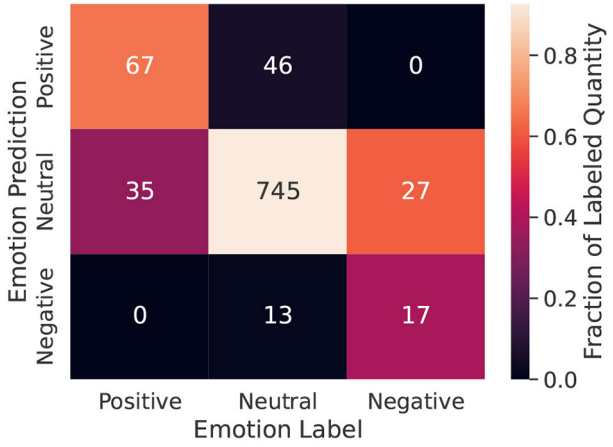

FIGURE 2
Confusion matrix for the best-performing Emotion model. Column sums: 102 Positive, 804 Neutral, and 44 Negative, 950 total examples.

6.4. Model performance

We compare the models from the grid searches in Section 6.3, and report the best model architectures, as determined by mean macro F1 score across five distinct holdout sets from cross-validation (Table 2). We use macro F1 score to select our top models because it provides a better estimate of performance for highly imbalanced classes, like the Negative class in Emotion classification.

This class imbalance is likely a driving factor behind the relatively poor macro F1 score for the Emotion task. Many of the ablation experiments that we tried were targeted to improve this performance, including augmentations and loss function. In Tables 3–5, we report results obtained when a single hyperparameter from the grid search is changed in the best model architecture (bolded) for each of the three tasks. We compare this series of BERT models to the best SVM model, and provide baseline human performance levels from the annotation procedure for comparison. Annotator accuracy/macro F1 is computed on the subset of labeled examples that the annotator contributed a label for, and individual annotator labels are compared to the adjudicated labels (considered truth values). In this way, the annotator predictions are *not independent* from the truth values, and as such, their accuracy and macro F1 scores are considered *upper bounds* on true annotator performance. We report the performance of two of the seven annotators, with Annotators A and B attaining the highest and lowest macro F1 scores among annotators, respectively.

As seen in the ablation tables, the BERT-based models outperform the SVM models. However, the performance of the top BERT-based model is not statistically better than many of the ablation experiments, given the reported standard deviation from the cross-validation test sets. For the Category and Structure tasks, the best-performing models outperform Annotator B's macro F1 score, and thus fall within the spread of human anotator performance. However, this was not the case for best model on the Emotion task, and we suggest a few possible reasons for the task being more difficult.

Figure 2 shows the confusion matrix as a heatmap for the best-performing Emotion model. The column sums represent the class distribution in our labeled dataset (also seen in Figure 1), with evident class imbalance. The biggest difficulty appears to be misclassifying Negative emotion posts as Neutral, which is understandable given the Negative class had the smallest number of examples across any class for any task in the 950 post training dataset. Despite these misclassifications, our auditing process (see Section 7) found 97% agreement for the Emotion labels after adjudication, perhaps indicating that many of these misclassified posts were "toss-ups" between Negative and Neutral, and that coding guidelines during initial labeling could have been clearer.

Figure 3 further examines the effect of each hyperparameter in the Emotion classification ablation studies, with a grid search conducted over 144 sets of model hyperparameters. For each hyperparameter value, the macro F1 score is averaged across all models trained with that value. We see, for example, that both `random` and `backtranslation` augmentation strategies outperform no augmentation on average. We also see that weighted sampling is very important for the Emotion task since `textual*`—the only Context hyperparameter setting *without* weighted sampling—performs worst.

Even though the macro F1 score of the best-performing Emotion model was lower than the score of Annotator B, possibly due to the class imbalance and coding issues outlined above, we considered the model's performance adequate for the Emotion task because the human annotator score is an upper bound (and considered a high target).

Since the focus of our work was to meet or exceed the performance of human annotators, enabling large-scale forum post classification, we didn't perform direct comparisons with alternative modeling approaches (e.g., different sets of features and model architectures). However, our model performance was on par with results from related works conducted on other MOOC forum datasets. For example, for category classification, Ntourmas et al. (2019) report an accuracy of 0.69 for an SVM model trained

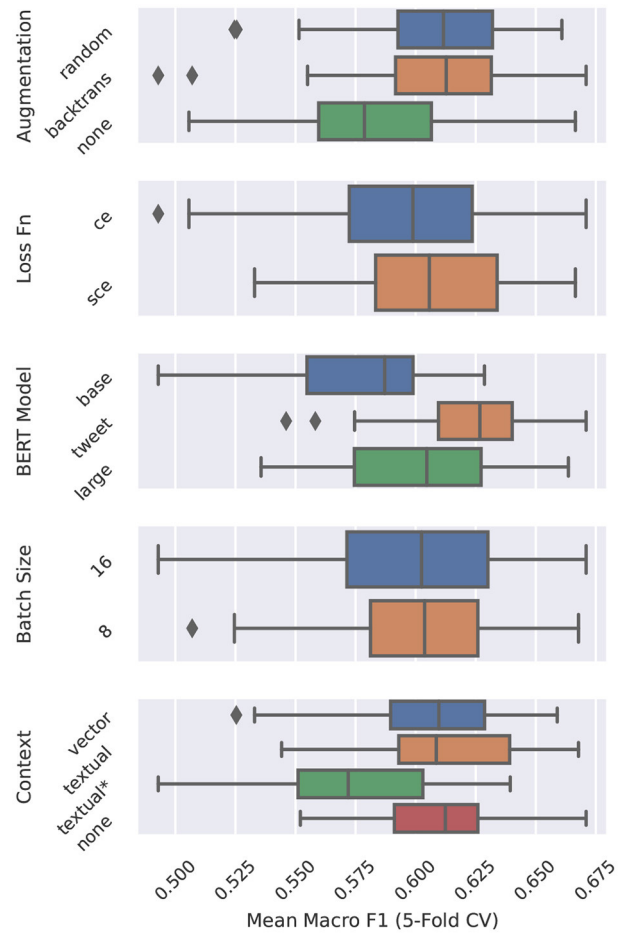

FIGURE 3
Aggregate performance of Emotion models while varying Augmentation, Loss Function, BERT Model, Batch Size, and Context. For Context, `textual*` indicates textual context format, without a weighted random sampler.

on an introductory Python course, and Ntourmas et al. (2021) achieve an accuracy of 0.64 on this course using the first two weeks of posts as training data for a decision tree classifier, and then evaluating the model on data from future weeks (three through six). Our best BERT-based model achieved an accuracy of 0.82 using 1% of the dataset, although we sample training data from all weeks of the MOOCs. For the emotion task, Clavié and Gal (2019) report accuracy of 89.78 for their EduBERT model on the StanfordMOOC sentiment dataset (where the original 7-point scale has been converted to binary with a score of 4 or above considered positive). Our best BERT model achieves 87.26 accuracy on a three-class formulation of the problem for our dataset. Finally, Sun et al. (2016) report an accuracy of 0.576 for a CRF applied to a 12-class dialogue act classification task on an edX MOOC dataset. We achieve an accuracy of 0.76 for our structure task, although our formulation has fewer classes (five).

Having achieved model performance that seemed on par with human annotators, we used the best models for each task to classify the remaining unlabeled data (approximately 81,000 posts). For each task, instead of training a new model on the full training set of 950 posts using the best hyperparameter setting (from Table 2), we created an ensemble of the five models trained during the five-fold cross-validation process associated with the best-performing setting. Ensembles have been shown to improve both accuracy and uncertainty calibration assuming enough diversity across ensemble members (Lakshminarayanan et al., 2017). Each of the five models trained during cross-validation was trained on a slightly different (though overlapping) subset of the training examples and had its non-pretrained weights initialized with different random seeds. Model predictions were then available for auditing (Step 8 in Section 3).

7. Auditing model outputs

Once the best models were identified, the model outputs were audited by team members for veracity. A team member randomly sampled 250 posts that had not been previously hand-coded. The team member blindly labeled the posts, then compared the hand-labeled posts with the best model label predictions (95% agreement for Category, 79% agreement for Structure, and 89% agreement for Emotion). Then the team-member labels and model labels were compared and adjudicated. After the adjudication session, there was a 96% agreement for Category, 87% for Structure, and 97% for Emotion. The improvements in Structure and Emotion were due to human error, which was detected in the adjudication process.

8. Model applications

With forum comments tagged and validated, we explored various forum characteristics and learner interactions to answer the research questions presented in Section 1. These questions serve as examples of how our AI models could be applied to understand structures and patterns in MOOC discussion forums. For RQ2, we investigated how the Category, Structure, and Emotion tasks are distributed in the forums and their relationship to one another across single posts as well as in comment threads. In RQ3 we asked how learners engaged with the forums, and if any group-level or course-level attributes emerged in their forum engagements. About a quarter of all posts were made by the Community TAs or instructors, and these posts were excluded from our analysis.

8.1. RQ2: use AI-assisted labeling to determine forum structures

In Figure 1, we presented the relative distribution of various classes under each task. As a next step, we analyzed how the different classes under each task overlapped with the classes from the other tasks. Our chosen metric is Jaccard similarity score, which for two sets A and B, is defined as $|A \cap B|/|A \cup B|$. The result is presented as a heatmap in Figure 4. As expected from the analysis of individual comments, there was a large overlap between the Content category and Neutral emotion (Jaccard score of 0.64). The Jaccard score of posts which fell in the Commentary category and Response/Comment structure was 0.44—the highest of any two classes among the structure and category tasks. The

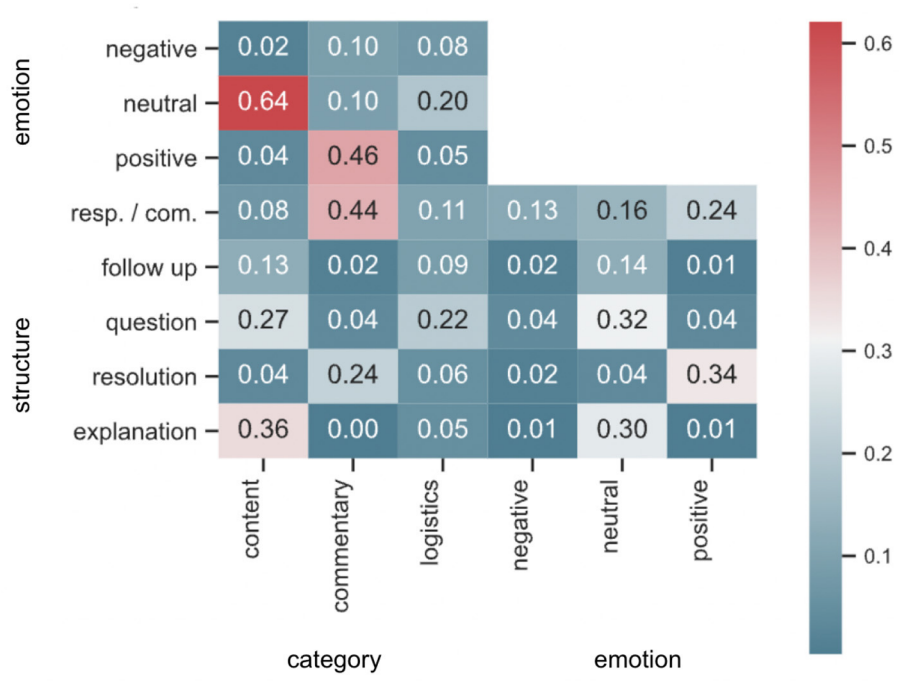

FIGURE 4
Jaccard similarity scores for AI-generated tags.

Commentary category had a higher Jaccard score with Positive emotion (0.46) than Negative emotion (0.10). The Structure tag of type Resolution had the Jaccard score of 0.34 with the Positive Emotion tag: the highest of all Emotion and Structure task pairs. Structure tags of types Question and Suggestion/Explanation were generally more Neutral emotionally, with scores 0.32 and 0.30, respectively. The highest score between the Structure and Category tasks were for posts categorized as Content with structure of type Suggestion/Explanation (0.36). The most common structure for the Logistics-category posts was Question (Jaccard score 0.22). Besides providing insight into the interrelationship among the tags, this analysis serves to further validate the output of our AI models by generating associations that are expected (e.g., the high score between Resolution structure and Positive emotion).

To understand the thread-level organization of forum posts, we aggregated posts for each thread, and determined how the Structure, Category, and Emotion tasks of each thread are distributed (e.g., a thread containing 2 Neutral posts, 1 Positive post, and 1 Negative post would have emotion fractions 0.5, 0.25, and 0.25, respectively). We binned these threads by the number of posts they contained, and presented the fraction of each task class as a function of the number of posts, averaged over the number of threads in a bin. We presented our analysis in Figure 5. There were fewer threads in the bins at the higher end of post counts, and the lines in the plots are more jagged.

We found that threads contained more general Commentary type structures as they grew in length, while the fraction of Questions kept falling. This downward trend of the Question fraction is expected: most threads had one question (and the follow-up questions had their own class). Explanation/Suggestion type structures showed a peak around 10 posts, followed by a slow descent. For the Emotion task, we found that the Neutral class occupied a smaller fraction in the longer threads, while the Negative and the Positive classes occupied a larger fraction. Similarly, the Commentary category occupied a larger fraction in the longer threads, at the cost of the Content- and Logistics-type categories. As in the previous analysis of Jaccard score among task classes, some of these trends (e.g., the Question fraction vs. the number of posts) returned intuitively expected answers, thus confirming the general validity of our AI-assisted labeling process.

8.2. RQ3: analysis of task associations with learner attributes

In this sub-section we analyze learner interactions with the forums using the AI-generated labels. We investigated two sub-questions: (a) How did the forum participants' demographic attributes and course attributes relate to the likelihood of their questions reaching resolution in a thread? (b) Was there any association between forum participants' attributes and whether they posted a comment in different classes of Structure, Category, and Emotion?

For each of these questions, we defined corresponding binary outcome variables indicating if there is at least one positive case, grouped at the learner level: i.e., to answer part (a) we looked at if a learner started at least one thread that reached a resolution. For part (b), the outcome variable was positive if a learner made at least one post of a certain type. We performed logistic regression with the

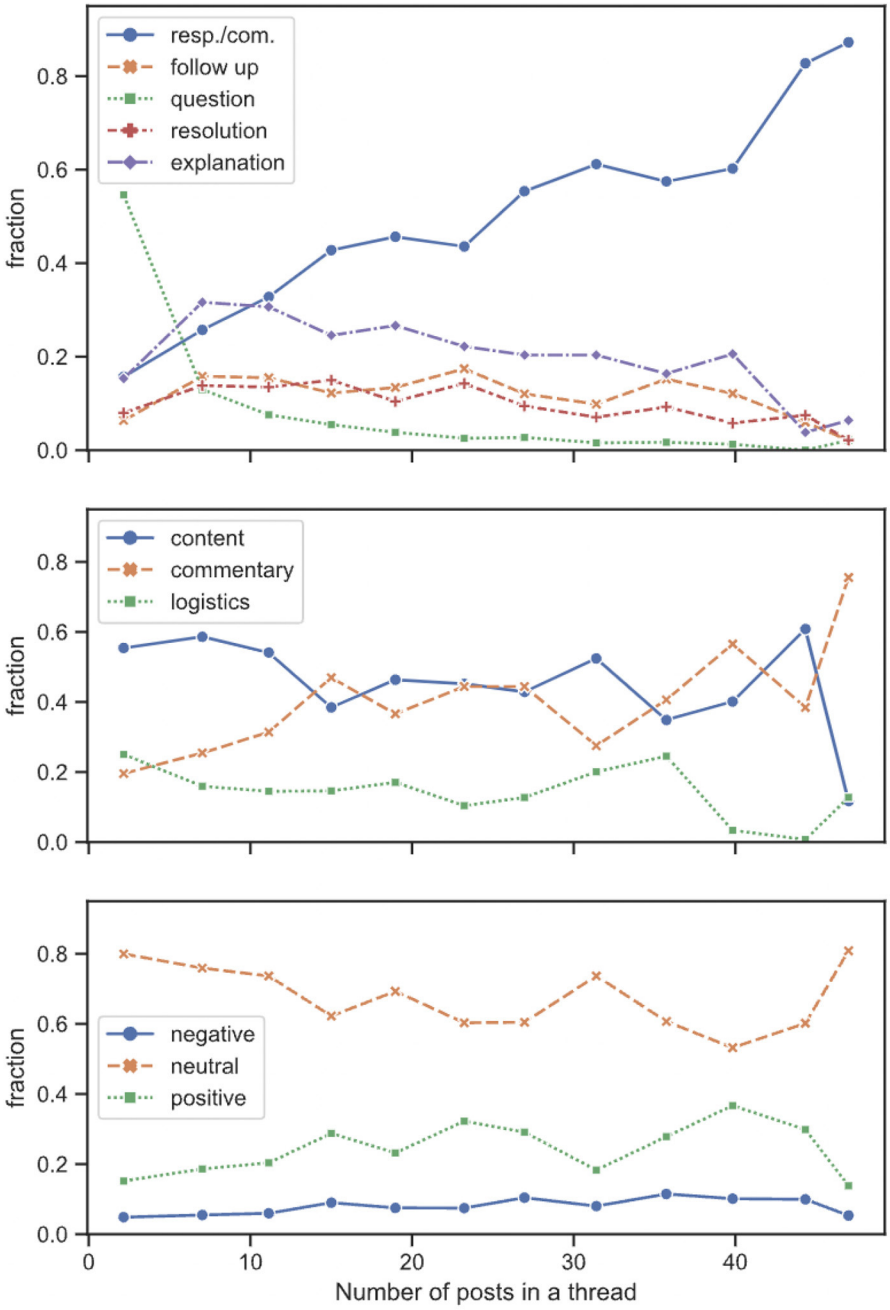

FIGURE 5
Fraction of task classes as functions of average number of posts in threads.

learner attributes and course attributes described below. In both cases, the likelihood would increase if a learner made more posts. To adjust for this, we first studied the number of posts authored by a learner (referred to as posts_frac), scaled by the maximum number of posts authored by any learner in our included courses, except the top 1% of posters. We found the top 1% of the forum posters have posted 31–181 posts, and we defined them as outliers, and set the maximum number of posts to 30. The re-scaling process helped us avoid over-weighting this variable when we included it as a covariate in the next analyses. We investigated how the posts_frac variable is associated with the learner attributes using a multiple linear regression model.

Within learner demographics, we considered their gender (male or female), age (below 22, 22-44, over 44 years old), education level (below bachelor's level, bachelors and above), and the economic status of their countries (high-income, upper- and lower-middle-income countries). For the learners' certification/verification status, there were three levels: those who were not verified (ineligible to earn certificates), those who were verified but did not earn certificates (`ver_not_cert`), and those

TABLE 6 Dependence of posts_frac on learner and course attributes. `lower_ci` and `upper_ci` refer to the lower and the upper bound of the 95% confidence intervals.

	coef	*p*-value	lower_ci	upper_ci
Intercept	0.0797	0.0000	0.0700	0.0893
gender_male	0.0017	0.6650	−0.0060	0.0093
age_22_to_44	0.0065	0.1024	−0.0013	0.0143
age_above_44	0.0398	0.0000	0.0296	0.0500
ver_cert	0.0702	0.0000	0.0637	0.0767
ver_not_cert	0.0140	0.0030	0.0048	0.0233
course_2	−0.0015	0.7468	−0.0104	0.0074

TABLE 7 Odds ratios, *p*-values, and the associated lower and upper bounds of 95% confidence intervals for the logistic regression model output for the learners whose questions are more likely to get resolved.

	OR	*p*-value	lower_ci	upper_ci
Intercept	0.2450	0.0000	0.2054	0.2922
gender_male	0.7080	0.0000	0.6025	0.8319
ver_cert	1.4905	0.0000	1.2882	1.7246
ver_not_cert	1.3107	0.0097	1.0677	1.6090
course_2	1.5086	0.0000	1.2552	1.8132
posts_frac	1e+02	0.0000	7e+01	2e+02

who were verified and did earn a certificate (`ver_cert`). We included the following course attributes: if the learner was enrolled in Python-1 or Python-2, and if the course ran during the COVID-19 pandemic or before it began (pandemic- or pre-pandemic-course). To make our regression analysis robust, we excluded posts where learner gender was recorded as "other," and if the learner was located in low-income countries, which reduced the learner count by less than 1%. However, such omission prompted by a need for greater statistical clarity reduces our understanding of the learners, and a larger learner sample size in a future study might mitigate this shortcoming.

In all of these regression models, the sample sizes were pretty large, ranging from $\sim$ 3000 to $\sim$ 8500 learners. When working with larger sample sizes, it is common to obtain small *p*-values. To avoid attributing statistical significance where it may not truly exist, we set the threshold *p*-value for significance at 0.0005. However, for the purpose of addressing part (b) of RQ3, we performed multiple tests and imposed a more stringent *p*-value threshold of 0.0001. The *p*-value thresholds are provided here as a heuristic to identify more important results, while the coefficients derived from our regression analyses indicate the effect size within these contexts.

In Table 6, we analyzed how posts_frac depended on the learner/course attributes, via a multiple linear regression model. In the final version of this model, we omitted demographic variables such as the education level, country income category, and whether the course ran during the pandemic or not, based on minimizing the *Akaike Information Score* (AIC). We found the biggest effect to be from the verified learners who went on to earn certificates: they posted about 7% more posts than those who were not verified (and did not earn certificates), after we controlled for other effects. The other significant result is from the learners $>$ 44 years old, who posted about 4% more on average. This result is important, as we find that these two groups (learners $>$ 44 years old and those who earned certificates) have significant results in our other models as well, even after we included the posts_frac as a covariate—meaning that these groups had a direct effect as well as an indirect effect (via posts_frac variable) on the outcome variables. The adjusted R-squared value for this model was $\sim$0.06.

To answer question (a), we first identified threads where the original posts were in the Content or Logistics category with Question-type structure. We aggregated such threads for each learner who started them and counted if at least one thread (for a learner) had a Resolution post, as determined by AI labeling. We did not explicitly take into account how many threads a learner started, but included posts_frac as a covariate. We presented the results from our logistic regression modeling for question (a) in Table 7. We chose the simplest model based on the pseudo-R-squared values, and in the final model we excluded learners' education level, country income category, and when the courses ran (i.e., during or before the pandemic). The overall analysis had a $p < 0.0001$. From our results it appeared that the learners who earned certificates and learners in course_2 had almost 50% higher likelihood of reaching a resolution in the threads they started, while male learners had $\sim$30% less odds of reaching a resolution, everything else being equal. The number of posts made by a learner was highly significant and the more a learner posted, the more likely it was that their thread would be resolved. In interpreting these odds ratios we need to remember that posts_frac is scaled to range from 0 to 1. Other factors (including the omitted ones) did not reach the level of significance we set earlier.

To answer question (b), we wanted to determine how learner and course attributes were associated with the likelihood that a learner made at least one post of a certain kind. To achieve this, we aggregated all posts made by a learner in a course, and defined the outcome variable as the likelihood of posting at least one post of a certain class (e.g., Question under the Structure task), as inferred from the AI-generated labels. We included posts_frac as a covariate in this case as well. We limited our analysis to six cases for brevity: Question and Suggestion/Explanation classes from the structure tasks, Positive and Negative classes from the emotion tasks, and Logistics and Content classes from the category tasks. We performed six separate logistic regression analyses: one for each of the above cases. In each of these cases, we aggregated all posts made by a user and noted if they made at least one post of that kind (e.g., negative emotion). For the independent variables, the final models excluded the education level, country income category, and course_pandemic, based on the same selection criteria we used in the last analysis.

In Table 8, we presented the odds ratios and the *p*-value resulting from these six models. With the *p*-value threshold of 0.0001, we found that compared to learners without verification, those who earned certificates were more likely to post at least one comment in all these categories except the structure type Suggestion/Explanation, and in the Content category. The trends were similar for verified learners without certificates, except they

TABLE 8 Odds ratios for included variables in logistic regression models.

	Structure question	Structure explain	Emotion negative	Emotion positive	Category logistics	Category content
Intercept	0.6588*	0.1548*	0.0649*	0.3101*	0.3097*	0.4102*
gender_male	0.8419	1.5624*	0.9093	0.6407*	0.9155	1.1508
age_22_to_44	0.6283*	1.3486*	0.9910	1.2133	0.7758*	1.3021*
age_above_44	0.6028*	1.3503	1.1330	1.0406	0.9493	1.2011
ver_cert	1.4768*	0.8846	1.3175	1.2597*	1.6228*	0.8900
ver_not_cert	1.4636*	0.6418*	1.5801*	1.3372*	1.6137*	0.7388*
course_2	1.2468	1.0231	1.0410	1.0154	0.9937	1.2279
posts_frac	9e+03*	7e+03*	3e+02*	2e+03*	2e+03*	5e+07*

Column names represent the dependent variables for individual regression models. Odds ratios with an asterisk (*) next to them indicate that they are statistically significant at $p \leq 0.0001$ level.

had significantly less likelihood in the Suggestion/Explanation (structure) or in the Content category. Both these groups (verified, with or without certificates) had a much higher likelihood of posting at least one Logistics-type post, and posts with Positive or Negative emotion, than all other groups. Male learners were found to be about 35% less likely to post at least one Positive-emotion post compared to female learners, but there was no significant difference between them when it came to Negative-emotion posts. Other things being equal, men were more likely to post at least one post of Suggestion/Explanation type, and less likely to ask questions. The last trend was similar for age groups 22–44 years old as well as for those 44 and above (compared to learners younger than 22 years old): the former two groups were less likely to post questions, and more likely to post suggestions/explanations. In addition, the learners in the 22–44 years old group were also more likely to post in the Content category, and less likely to post in the Logistics category. As expected, posts_frac strongly increased the likelihood in all six cases. Additionally, we noticed a couple of inversely related trends: the likelihoods of posting a Question and a Suggestion/Explanation (under the Structure tasks) were inversely related, and so were the likelihoods for the Content and the Logistics categories. In Section 9.2, we discuss possible explanations and implications of these findings.

9. Discussion

9.1. Implications of AI-assisted labeling

We successfully applied the proposed AI-assisted labeling process to generate a fully labeled dataset while only requiring our team to annotate approximately 1% of the posts. The AI-generated labels were found to be reliable and formed the foundation of analyses that spanned the category, structure, and emotion dimensions as well as learner demographics and course attributes.

However, the ability to directly apply the models trained on the Python MOOC dataset to other MOOCs is unknown [although Bakharia (2016) and Ntourmas et al. (2019) report poor cross-domain performance in their settings]. Future work could explore the impact of classification task (Category, Structure, or Emotion) on transferability to other courses.

Even if the models trained on the edX Python courses don't transfer well, we believe this general approach can be extended to non-CS courses on any platform, with appropriate training and testing. There was little direct impact of content on model accuracy, except its homogeneity, and the context added to the posts are replicable from most other forums, making this process platform-independent. One promising future direction of our work would be to evaluate this method for non-CS online courses running on different platforms.

Although we achieved a substantial reduction in annotation effort, making application to new courses feasible, the reduction achieved will likely vary by dataset. Future directions to reduce the burden further include: using active learning (Ren et al., 2021) to identify the most impactful/useful next posts to annotate, using "data programming" techniques (Ratner et al., 2016, 2017) to programmatically generate labels using heuristics, trying other techniques for dealing with class imbalance such as explicitly oversampling likely examples from the minority classes for annotation or applying techniques like SMOTE (Chawla et al., 2002) during model training, using semi-supervised techniques such as self-training as a way to leverage the annotated examples more effectively by pseudo-labeling unlabeled examples, and exploring zero-shot or few-shot learning with large language models (LLMs) such as GPT-3 (Brown et al., 2020). Annotating whole threads at a time (vs. individual posts sampled from disjoint threads) might potentially make the annotation task easier and more reliable (in terms of inter-annotator agreement). Analogously, graph neural networks (Wu et al., 2020) could potentially increase consistency of post labels within threads as well as boost accuracy for the Structure task.

Lastly, recent advances in instruction-tuned LLMs (Ouyang et al., 2022) such as ChatGPT (OpenAI, 2023) suggest the possibility of using an LLM-based agent to draft responses to questions brought up in the forum, which could help improve responsiveness to learners and reduce the burden on course staff. Though a promising future direction, several LLM challenges (OpenAI, 2023; Touvron et al., 2023) would need to be addressed, including hallucination (generating content that is untruthful), producing biased or toxic content, and dealing with new or changing information not present in the training set.

9.2. Implications of our findings from model applications

The improvement in coding performance from using the AI-assisted labeling made it possible to extensively analyze learners' engagement in the discussion forums, and this knowledge could be helpful in multiple ways. When re-designing a course, instructors can preemptively take action to identify and address some recurring issues, such as commonly occurring logistics problems or content-based questions. During an ongoing course with a high volume of forum posts, instructors/TAs may choose to prioritize some posts over others (e.g., the logistics-related posts or those with negative emotions).

Our analysis of ordinary forum participants (i.e., excluding the outliers) show that learners who earned certificates post more, as well as the learners older than 44 years. These results corroborate earlier research findings which showed a correlation between forum posts and course performance (Wen et al., 2014; Houston et al., 2017; Wise and Cui, 2018). On the other hand, we did not observe a strong effect of gender once we controlled for other factors. By bringing together these demographic attributes and course attributes along with the AI-assisted annotations, the present study furthers our understanding of MOOC forum dynamics.

The differences among groups in how learners participated in the forum are often intuitive. For example, verified learners/certificate earners asked more questions in the Logistics category. In MOOCs, these groups of learners often progress more quickly through the course than non-verified learners, and thus these groups are more likely to face issues related to logistics before others. Once these issues are brought to notice, they may get fixed or some workaround is posted in the forum, thus reducing the need for posting Logistics category posts for those attempting the same course component later. On the other hand, the non-verified learners may face more challenges in the content area, and may not be able to find answers on the forums, as the verified learners may not have experienced the same difficulty with content, or they have moved on to a different section of the course. The finding that the verified learners post more comments with positive or negative emotions may be a clue that they are more engaged with the course and more invested in their outcomes, and the joy and frustration of learning they experience are more pronounced than the non-verified learners. This observation was also seen by Wen et al. (2014), who found that both strong positive and negative sentiments were associated with high completion rates for active learners.

From our analysis, we found that verified learners were less likely to post suggestions and explanations. It may be argued that they were better prepared than unverified learners to offer suggestions, but less willing. However, the more likely explanation could be that they may not encounter many other learners seeking help at a point when they are engaged with specific content. If this is indeed true, a course re-design where more advanced learners were encouraged to share their challenges with others is likely to benefit everyone.

From the same analysis, we found systemic differences in how men and women use the forums. For instance, fewer men are likely to post comments with positive emotion than women (while no such difference exists for negative emotion). Whether such differences make women participants feel less welcome on online forums is not immediately obvious, but is worth exploring further.

While we focused on a handful of applications of AI-assisted labeling, we can extend the same tools to analyze how the forum interactions of individual users change with course progression. We can study the formation of learner groups within forums and the intragroup interaction. This same process could be extended to other courses where a tasks-based description of forums is meaningful, and to other learning platforms where we can find similar context to what we provided to our AI models. We believe that learners who participate in online forums are more motivated to do well in the courses, and the AI-assisted labeling methods to analyze forum participation can provide insight with less human labor, thus enabling course designers and instructors to serve their learners better.

10. Conclusion

This paper sought to explore how learners are engaging with discussion forums in MOOCs to gain support and help each other through the course. We were able to develop a BERT-based model architecture and training approach that supports fine-tuning pretrained models to custom forum post classification tasks using only a small amount of manually generated annotations. We then could use this model to categorize discussion posts to enable an analysis of the data. We showed that different groups of learners often differ significantly in how they interact with the forums: whether they discuss course content or the logistics, the sentiment they show in their posts, and whether they ask questions or offer suggestions. We found that learner attributes of those who start comments threads have strong association with the likelihood that a thread reaches resolution. Future research can build on this work, exploring learner engagement in forums and the impacts of targeted interventions.

Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: the datasets are not readily available because of potential inclusion of sensitive personal information such as username or email ids in the forum comments, which are difficult to systematically remove. Research access to the data may be available with appropriate authorization and approval from concerned institutional review boards. Requests to access these datasets should be directed to AB, anabell@mit.edu.

Ethics statement

Ethical approval was not required for the study involving human data in accordance with the local legislation and institutional requirements. Written informed consent to participate in this study was not required in accordance with the national legislation and the institutional requirements.

Author contributions

MY, ARo, MP, CC, and SM contributed to the conception and design of the study. AB provided access to the MOOC datasets. MY, ARo, MP, CC, KQ, and ARu annotated data. CC and KQ designed, trained, and analyzed AI models. MP audited model outputs. ARo performed data analysis of model outputs and learner data. MY, ARo, MP, and KQ wrote sections of the manuscript. All authors approved the submitted version.

Funding

Research was sponsored by the United States Air Force Research Laboratory and the Department of the Air Force Artificial Intelligence Accelerator and was accomplished under Cooperative Agreement Number FA8750-19-2-1000.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Author disclaimer

The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Department of the Air Force or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

Agrawal, A., Venkatraman, J., Leonard, S., and Paepcke, A. (2015). *YouEDU: Addressing Confusion in MOOC Discussion Forums by Recommending Instructional Video Clips*. International Educational Data Mining Society.

Ahmad, M., Junus, K., and Santoso, H. B. (2022). Automatic content analysis of asynchronous discussion forum transcripts: a systematic literature review. *Educ. Inform. Technol.* 27, 11355–11410. doi: 10.1007/s10639-022-11065-w

Almatrafi, O., and Johri, A. (2018). Systematic review of discussion forums in massive open online courses (MOOCs). *IEEE Trans. Learn. Technol.* 12, 413–428. doi: 10.1109/TLT.2018.2859304

Alrajhi, L., Alharbi, K., and Cristea, A. I. (2020). "A multidimensional deep learner model of urgent instructor intervention need in MOOC forum posts," in *Intelligent Tutoring Systems: 16th International Conference, ITS 2020* (Athens: Springer), 226–236.

Bakharia, A. (2016). "Towards cross-domain MOOC forum post classification," in *Proceedings of the Third 2016 ACM Conference on Learning@ Scale* (New York, NY), 253–256. Available online at: https://dl.acm.org/doi/abs/10.1145/2876034.2893427

Barbieri, F., Camacho-Collados, J., Espinosa-Anke, L., and Neves, L. (2020). "TweetEval: unified benchmark and comparative evaluation for Tweet classification," in *Proceedings of Findings of EMNLP* (Stroudsburg, PA). Available online at: https://aclanthology.org/2020.findings-emnlp.148/

Boroujeni, M. S., Hecking, T., Hoppe, H. U., and Dillenbourg, P. (2017). "Dynamics of MOOC discussion forums," in *Proceedings of the Seventh International Learning Analytics & Knowledge Conference* (New York, NY), 128–137. Available online at: https://dl.acm.org/doi/abs/10.1145/3027385.3027391

Brinton, C. G., Chiang, M., Jain, S., Lam, H., Liu, Z., and Wong, F. M. F. (2014). Learning about social learning in MOOCs: from statistical analysis to generative model. *IEEE Trans. Learn. Technol.* 7, 346–359. doi: 10.1109/TLT.2014.2337900

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., et al. (2020). "Language models are few-shot learners," in *Advances in Neural Information Processing Systems 33*, eds H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Red Hook, NY: Curran Associates, Inc.), 1877–1901.

Capuano, N., Caballé, S., Conesa, J., and Greco, A. (2021). Attention-based hierarchical recurrent neural networks for MOOC forum posts analysis. *J. Ambient Intell. Hum. Comput.* 12, 9977–9989. doi: 10.1007/s12652-020-02747-9

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *J. artif. Intell. Res.* 16, 321–357. doi: 10.1613/jair.953

Chen, J., Feng, J., Sun, X., and Liu, Y. (2019). Co-training semi-supervised deep learning for sentiment classification of MOOC forum posts. *Symmetry* 12, 8. doi: 10.3390/sym12010008

Clavié, B., and Gal, K. (2019). EduBERT: pretrained deep language models for learning analytics. *arXiv preprint arXiv:1912.00690*. doi: 10.48550/arXiv.1912.00690

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*. doi: 10.48550/arXiv.1810.04805

Ezen-Can, A., Boyer, K. E., Kellogg, S., and Booth, S. (2015). "Unsupervised modeling for understanding MOOC discussion forums: a learning analytics approach," in *Proceedings of the Fifth International Conference on Learning Analytics and Knowledge* (New York, NY), 146–150. Available online at: https://dl.acm.org/doi/abs/10.1145/2723576.2723589

Fisher, R., Simmons, R., and Malin-Mayor, C. (2015). "Weakly supervised learning of dialogue structure in MOOC forum threads," in *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)* (Miami, FL: IEEE), 624–627. Available online at: https://ieeexplore.ieee.org/abstract/document/7424387

Galikyan, I., Admiraal, W., and Kester, L. (2021). MOOC discussion forums: the interplay of the cognitive and the social. *Comput. Educ.* 165, 104133. doi: 10.1016/j.compedu.2021.104133

Gillani, N., and Eynon, R. (2014). Communication patterns in massively open online courses. *Internet Higher Educ.* 23, 18–26. doi: 10.1016/j.iheduc.2014.05.004

Guo, S. X., Sun, X., Wang, S. X., Gao, Y., and Feng, J. (2019). Attention-based character-word hybrid neural networks with semantic and structural information for identifying of urgent posts in MOOC discussion forums. *IEEE Access* 7, 120522–120532. doi: 10.1109/ACCESS.2019.2929211

Houston, S. L., Brady, K., Narasimham, G., and Fisher, D. (2017). "Pass the idea please: the relationship between network position, direct engagement, and course performance in MOOCs," in *Proceedings of the Fourth 2017 ACM Conference on Learning@ Scale* (New York, NY), 295–298. Available online at: https://dl.acm.org/doi/abs/10.1145/3051457.3054008

Huang, J., Dasgupta, A., Ghosh, A., Manning, J., and Sanders, M. (2014). "Superposter behavior in MOOC forums," in *Proceedings of the First ACM Conference on Learning@ Scale Conference* (New York, NY), 117–126. Available online at: https://dl.acm.org/doi/abs/10.1145/2556325.2566249

John, C., and Meinel, C. (2020). "Learning behavior of men and women in MOOC discussion forums–a case study," in *2020 IEEE Global Engineering Education Conference (EDUCON)* (Porto: IEEE), 300–307. Available online at: https://ieeexplore.ieee.org/abstract/document/9125322

Joksimović, S., Jovanović, J., Kovanović, V., Gašević, D., Milikić, N., Zouaq, A., and Van Staalduinen, J. P. (2019). Comprehensive analysis of discussion forum participation: from speech acts to discussion dynamics and course outcomes. *IEEE Trans. Learn. Technol.* 13, 38–51. doi: 10.1109/TLT.2019.2916808

Kim, J., and Kang, J.-H. (2014). Towards identifying unresolved discussions in student online forums. *Appl. Intell.* 40, 601–612. doi: 10.1007/s10489-013-0481-1

Krippendorff, K. (1970). Estimating the reliability, systematic error and random error of interval data. *Educ. Psychol. Meas.* 30, 61–70.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2017). "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Advances in Neural Information Processing Systems 30*, eds I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Red Hook, NY: Curran Associates, Inc.), 6402–6413.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Lee, J., Soleimani, F., Hosmer, J. IV, Soylu, M. Y., Finkelberg, R., and Chatterjee, S. (2022). Predicting cognitive presence in at-scale online learning: MOOC and for-credit online course environments. *Online Learn.* 26, 58–79. doi: 10.24059/olj.v26i1.3060

Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., et al. (2022). A survey on text classification: from traditional to deep learning. *ACM Trans. Intell. Syst. Technol.* 13, 1–41. doi: 10.1145/3495162

Li, X., Zhang, H., Ouyang, Y., Zhang, X., and Rong, W. (2019). "A shallow bert-cnn model for sentiment analysis on MOOCs comments," in *2019 IEEE International Conference on Engineering, Technology and Education (TALE)* (Yogyakarta: IEEE), 1–6. Available online at: https://ieeexplore.ieee.org/abstract/document/9225993

Ma, E. (2019). *NLP Augmentation.* Available online at: https://github.com/makcedward/nlpaug

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*. doi: 10.48550/arXiv.1301.3781

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems 26*, eds C. J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Red Hook, NY: Curran Associates, Inc.), 3111–3119.

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., and Gao, J. (2021). Deep learning–based text classification: a comprehensive review. *ACM Comput. Surv.* 54, 1–40. doi: 10.1145/3439726

Moreno-Marcos, P. M., Alario-Hoyos, C., Muñoz-Merino, P. J., Estevez-Ayres, I., and Kloos, C. D. (2018). A learning analytics methodology for understanding social interactions in MOOCs. *IEEE Trans. Learn. Technol.* 12, 442–455. doi: 10.1109/TLT.2018.2883419

Nadeau, C., and Bengio, Y. (1999). "Inference for the generalization error," in *Advances in Neural Information Processing Systems*, Vol. 12, eds S. Solla, T. Leen, and K. Müller (Cambridge, MA: MIT Press), 307–313.

Ng, N., Yee, K., Baevski, A., Ott, M., Auli, M., and Edunov, S. (2019). Facebook fair's WMT19 news translation task submission. *arXiv preprint arXiv:1907.06616*. doi: 10.48550/arXiv.1907.06616

Niu, T., and Bansal, M. (2018). "Adversarial over-sensitivity and over-stability strategies for dialogue models," in *The SIGNLL Conference on Computational Natural Language Learning (CoNLL)* (Stroudsburg, PA). Available online at: https://aclanthology.org/K18-1047/

Ntourmas, A., Avouris, N., Daskalaki, S., and Dimitriadis, Y. (2019). "Comparative study of two different MOOC forums posts classifiers: analysis and generalizability issues," in *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)* (Patras: IEEE), 1–8. Available online at: https://ieeexplore.ieee.org/abstract/document/8900682

Ntourmas, A., Dimitriadis, Y., Daskalaki, S., and Avouris, N. (2021). "Classification of discussions in MOOC forums: an incremental modeling approach," in *Proceedings of the Eighth ACM Conference on Learning@ Scale* (New York, NY), 183–194. Available online at: https://dl.acm.org/doi/abs/10.1145/3430895.3460137

OpenAI (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*. doi: 10.48550/arXiv.2303.08774

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., et al. (2022). "Training language models to follow instructions with human feedback," in *Advances in Neural Information Processing Systems 35*, eds S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Red Hook, NY: Curran Associates, Inc.), 27730–27744.

Pang, B., and Lee, L. (2008). Opinion mining and sentiment analysis. *Found. Trends Inform. Retrieval* 2, 1–135. doi: 10.1561/1500000011

Pennington, J., Socher, R., and Manning, C. D. (2014). "GLOVE: global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Stroudsburg, PA), 1532–1543. Available online at: https://aclanthology.org/D14-1162/

Poquet, O., and Dawson, S. (2015). "Analysis of MOOC forum participation," in *Ascilite15 - Conference of the Australasian Society for Computers in Learning in Tertiary Education* (Perth), 224. Available online at: http://www.2015conference.ascilite.org/wp-content/uploads/2015/11/ascilite-2015-proceedings.pdf

Poquet, O., Dowell, N., Brooks, C., and Dawson, S. (2018). "Are MOOC forums changing?" in *Proceedings of the 8th International Conference on learning Analytics and Knowledge* (New York, NY), 340–349. Available online at: https://dl.acm.org/doi/abs/10.1145/3170358.3170416

Prechelt, L. (2012). *Early Stopping — But When?*. Berlin; Heidelberg: Springer.

Ratner, A., Bach, S. H., Ehrenberg, H., Fries, J., Wu, S., and Ré, C. (2017). "Snorkel: Rapid training data creation with weak supervision," in *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases* (Rio de Janeiro: NIH Public Access), 269. Available online at: https://dl.acm.org/doi/10.14778/3157794.3157797; https://dl.acm.org/toc/pvldb/2017/11/3

Ratner, A. J., De Sa, C. M., Wu, S., Selsam, D., and Ré, C. (2016). "Data programming: Creating large training sets, quickly," in *Advances in Neural Information Processing Systems 29*, eds D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (Red Hook, NY: Curran Associates, Inc.), 3567–3575.

Ren, P., Xiao, Y., Chang, X., Huang, P.-Y., Li, Z., Gupta, B. B., et al. (2021). A survey of deep active learning. *ACM Comput. Surv.* 54, 1–40. doi: 10.1145/3472291

Rosenthal, S., Farra, N., and Nakov, P. (2017). "SemEval-2017 task 4: sentiment analysis in Twitter," in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)* (Stroudsburg, PA), 502–518. Available online at: https://aclanthology.org/S17-2088/

Roy, A., Yee, M., Perdue, M., Stein, J., Bell, A., Carter, R., et al. (2022). "How COVID-19 affected computer science MOOC learner behavior and achievements: a demographic study," in *Proceedings of the Ninth ACM Conference on Learning @ Scale, L@S '22* (New York, NY: Association for Computing Machinery), 345–349.

Schmid, P. (2021). *Huggingface BERT-tweet-eval-emotion*. Available online at: https://huggingface.co/philschmid/BERT-tweet-eval-emotion

Schütze, H., Manning, C. D., and Raghavan, P. (2008). *Introduction to Information Retrieval*, Vol. 39. Cambridge: Cambridge University Press.

Sennrich, R., Haddow, B., and Birch, A. (2015). Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*. doi: 10.48550/arXiv.1511.06709

Sha, L., Rakovic, M., Li, Y., Whitelock-Wainwright, A., Carroll, D., Gaševic, D., et al. (2021). *Which Hammer Should I Use? A Systematic Evaluation of Approaches for Classifying Educational Forum Posts*. International Educational Data Mining Society.

Sun, C., Li, S.-w., and Lin, L. (2016). "Thread structure prediction for MOOC discussion forum," in *Social Computing: Second International Conference of Young Computer Scientists, Engineers and Educators, ICYCSEE 2016* (Harbin: Springer), 92–101.

Sun, X., Guo, S., Gao, Y., Zhang, J., Xiao, X., and Feng, J. (2019). "Identification of urgent posts in MOOC discussion forums using an improved rCNN," in *2019 IEEE World Conference on Engineering Education (EDUNINE)* (Lima: IEEE), 1–5. Available online at: https://ieeexplore.ieee.org/abstract/document/8875845

Swinnerton, B., Hotchkiss, S., and Morris, N. P. (2017). Comments in MOOCs: who is doing the talking and does it help? *J. Comput. Assist. Learn.* 33, 51–64. doi: 10.1111/jcal.12165

Topping, K. J. (2005). Trends in peer learning. *Educ. Psychol.* 25, 631–645. doi: 10.1080/01443410500345172

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., et al. (2023). Llama: open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*. doi: 10.48550/arXiv.2302.13971

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *Advances in Neural Information Processing Systems 30*, eds I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Red Hook, NY: Curran Associates, Inc.), 5998–6008.

Wen, M., Yang, D., and Rose, C. (2014). "Sentiment analysis in MOOC discussion forums: What does it tell us?" in *Educational Data Mining 2014*, eds J. Stamper, Z. Pardos, M. Mavrikis, and B. M. McLaren (London: Citeseer), 130–137.

Wise, A. F., and Cui, Y. (2018). Learning communities in the crowd: characteristics of content related interactions and social relationships in MOOC discussion forums. *Comput. Educ.* 122, 221–242. doi: 10.1016/j.compedu.2018.03.021

Wise, A. F., Cui, Y., Jin, W., and Vytasek, J. (2017). Mining for gold: identifying content-related MOOC discussion threads across domains through linguistic modeling. *Internet Higher Educ.* 32, 11–28. doi: 10.1016/j.iheduc.2016.08.001

Wong, J.-S., Pursel, B., Divinsky, A., and Jansen, B. J. (2015). "An analysis of MOOC discussion forum interactions from the most active users," in *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction* (Heidelberg: Springer), 452–457. Available online at: https://link.springer.com/chapter/10.1007/978-3-319-16268-3_58#chapter-info

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., and Philip, S. Y. (2020). A comprehensive survey on graph neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* 32, 4–24. doi: 10.1109/TNNLS.2020.2978386

Yamarik, S. (2007). Does cooperative learning improve student learning outcomes? *J. Econ. Educ.* 38, 259–277. doi: 10.3200/JECE.38.3.259-277

Yang, D., Wen, M., Howley, I., Kraut, R., and Rose, C. (2015). "Exploring the effect of confusion in discussion forums of massive open online courses," in *Proceedings of the Second (2015) ACM Conference on Learning@ Scale* (New York, NY), 121–130. Available online at: https://dl.acm.org/doi/abs/10.1145/2724660.2724677

Yee, M., Roy, A., Stein, J., Perdue, M., Bell, A., Carter, R., et al. (2022). "The relationship between COVID-19 severity and computer science MOOC learner achievement: a preliminary analysis," in *Proceedings of the Ninth ACM Conference on Learning @ Scale, L@S '22* (New York, NY: Association for Computing Machinery), 431–435.

Zhang, L., Wang, S., and Liu, B. (2018). Deep learning for sentiment analysis: a survey. *Wiley Interdiscipl. Rev. Data Mining Knowledge Discov.* 8, e1253. doi: 10.1002/widm.1253

Zou, W., Hu, X., Pan, Z., Li, C., Cai, Y., and Liu, M. (2021). Exploring the relationship between social presence and learners' prestige in MOOC discussion forums using automated content analysis and social network analysis. *Comput. Hum. Behav.* 115, 106582. doi: 10.1016/j.chb.2020.106582