Assessing Algorithmic Fairness in Automatic Classifiers of Educational Forum Posts

Lele Sha[1,3], Mladen Rakovic[1,3], Alexander Whitelock-Wainwright[1,2], David Carroll[1,2], Victoria M. Yew, Dragan Gasevic[1,3], and Guanliang Chen*[1,3]

[1] Centre for Learning Analytics at Monash, Faculty of Information Technology, Monash University, Australia
[2] Portfolio of the Deputy Vice-Chancellor Education, Monash University, Australia
[3] Department of Data Science and Artificial Intelligence, Faculty of Information Technology, Monash University, Australia
{lele.sha1,mladen.rakovic,alex.wainwright,david.carroll, dragan.gasevic,guanliang.chen}@monash.edu, victoriayew@gmail.com

Abstract. Automatic classifiers of educational forum posts are essential in helping instructors effectively implement their teaching practices and thus have been widely investigated. However, existing studies mostly stressed the *accuracy* of a classifier, while the *fairness* of the classifier remains largely unexplored, i.e., whether the posts generated by a group of students are more likely to be correctly labeled than those generated by other groups of students. Undoubtedly, any unfairness based on student performance, sex, or other subjective views can have a detrimental effect on a student's learning experience and performance. Therefore, this study aimed to assess the algorithmic fairness of six popular models used in building automatic classifiers of educational forum posts. Here, we measured the algorithmic fairness displayed (i) between students of different sex (female vs. male) and (ii) between students of different first languages (English-as-first-language speakers vs. English-as-second-language speakers). Besides, we investigated whether a classifier's fairness could be enhanced by applying data sampling techniques. Our results demonstrated that: 1) traditional Machine Learning models slightly outperformed up-to-date Deep Learning models in delivering fair predictions; 2) students of different first languages faced more unfair predictions than students of different sex, and most of the classifiers tended to favor English-as-first-language students; and 3) with equal numbers of posts generated by different groups of students in the training data, the fairness of a classifier could be greatly enhanced.

Keywords: Educational Forum Post · Text Classification · Algorithmic Fairness

1 Introduction

Students in online courses are afforded opportunities to earn university credentials while learning remotely in a self-directed manner. Unlike traditional

* Corresponding author.

in-person classes, regular communication between course instructors and individual students in online courses is often sparse [26, 27], despite the documented benefits of instructor's presence [23]. For this reason, many students in online classes feel deprived of necessary guidance and support [22]. This often hinders students' satisfaction and learning performance in an online course [24]. Communication among students themselves in an online learning environment has been considered a remedy for an instructor's absence [47, 49]. A sense of community that students build in online discussion forums boosts their engagement and satisfaction in a course [36, 43, 49]. Importantly, productive forum discussions that unfold throughout a semester have been shown to benefit learning gains [44]. It is, therefore, critical for online students to create course-relevant discussion posts. To this end, educators need to continuously monitor discussion boards, identify posts that require instructors' urgent attention (e.g., posts asking questions related to the course learning content) and provide timely support to students. This is, however, a time-consuming task in many online courses, given an abundant number of posts typically created on discussion boards.

To address this challenge, educational researchers have developed a number of classifiers to automatically identify content-relevant and content-irrelevant discussion posts (whether the post content is related to knowledge taught in a course). To our knowledge, both traditional Machine Learning (ML) models, e.g., Random Forests, and up-to-date Deep Learning (DL) models, e.g., Long Short-Term Memory Neural Network (LSTM), have been exploited for this classification task [2, 6, 9, 13, 17, 18, 21, 42, 50, 52, 54]). While many of these models have demonstrated attractive classification accuracy, none of them has reported classification performance evaluated relative to different demographic groups in the student sample. Given the raising concerns about algorithmic unfairness of predictive models in educational research [19] and widely documented discrepancies in retention between female and male students, particularly in STEM courses [16, 40], and cognitive and social barriers that many English-as-second-language speakers face when communicating about topics taught in English [20, 32, 34, 39], we posit that the development of more inclusive educational technologies grounded in fair classification models that perform equally well across all groups of students, including their sex and first-language backgrounds, should be an important next step in the educational research agenda.

With this in mind, this study set out to assess not only the accuracy but also the fairness of popular models used to construct automatic classifiers of educational forum posts, including four ML models and two DL models. In particular, the fairness of these models was measured by distinguishing students of different sex and first-language backgrounds. Through extensive evaluations, we demonstrated that classifiers of educational forum posts were prone to algorithmic unfairness in classifying posts created by students of different sex and first languages. To address model unfairness, we explored the viability of equal sampling of the observed demographic groups in the model training process. Our results indicated that most of the models improved their fairness, suggesting

that equal sampling can be an important step in the future development of fair classifiers of educational forum posts.

2 Related Work

Educational Forum Post Classification. Research efforts in predictive modeling for educational forum posts have generally relied upon traditional ML or DL models. Among traditional ML models, researchers have frequently used Random Forests [2, 5, 18, 30, 35, 37], Support Vector Machine (SVM) [5, 13, 17, 28, 35, 38, 42, 53], Logistic Regression [1, 2, 35, 52, 56, 58], and Naïve Bayes [4, 5, 35]. It should be noted that these models were often based on features engineered by experts. For instance, when applying Random Forests to classify forum posts by levels of cognitive presence, researchers in [18] designed 87 different features as proxies for cognitive presence, e.g., the length of a post, the semantic similarity between posts, number of replies a post received, and frequency of words indicative of different psychological processes. All together, these features enabled the Random Forest classifier to achieve a Cohen's κ score of 0.72. As another example, Cui et al. [13] engineered post features including unigrams and bigrams of a post text, the number of views and votes a post attracted, part-of-speech tags; they used these features to create an SVM classifier to distinguish content-relevant posts from content-irrelevant ones to assist instructors to identify posts that require urgent attention in MOOC discussion forums. With the SVM classifier, about 86% of forum posts were accurately identified.

In recent years, driven by the great success achieved by DL models in tackling various prediction tasks, a growing number of researchers has opted for DL models to classify educational forum posts [3, 9, 11, 12, 21, 50, 54]. Compared to traditional ML models, DL models do not require domain experts to carefully design features as input. Instead, DL models can take the raw text of a post as input and make use of the powerful affordances of deep neural networks to implicitly capture features that are important to correctly classify a post. For example, one of the pioneering studies that applied DL models was reported in [50]. Specifically, the authors in [50] developed ConvL, a DL classifier that identifies different levels of urgency, confusion, and sentiment in educational forum posts. The classifier development involved two important steps. Firstly, the researchers applied CNN to derive contextual features related to a post and, secondly, used LSTM to capture sequential relationships between these features for classification. Evaluated on the dataset with more than 30 thousand educational forum posts, ConvL achieved accuracy between 81% and 87%. Other relevant studies that applied DL models for post classifications tasks, typically, relied on CNN, LSTM, or variants of these two models [54]. Moreover, a recent study reported in [12] demonstrated that, even when the size of annotated data is insufficient to support the training of DL models, pre-trained language models, e.g., BERT [14], could be exploited to empower those DL models for post classification. Specifically, the researchers in [12] showed that, by simply coupling only one classification neural network layer on top of the output of BERT, the classifica-

tion accuracy could be boosted up to to 92%. Though researchers have achieved great advances in constructing accurate classifiers of educational forum posts, it remains largely unknown whether these classifiers generate equally accurate predictions to different groups of students. To our knowledge, our study is the first to investigate the problem of fairness in constructing post classifiers. To this end, we assessed the capability of a total of six different models in generating both accurate and fair classification results for different groups of students, which were created as per their sex (female and male) and first-language background (English-as-first-language vs English-as-second-language speakers).

Fairness-aware Machine Learning Models in Education. As witnessed by the establishment of the ACM Conference on Fairness, Accountability, and Transparency in 2018, one of the recent foci in the broader ML community is to assess algorithmic unfairness of different intelligent systems and investigate approaches for alleviating the negative impacts brought by such algorithmic unfairness. In the educational research field, a few studies have been carried out to investigate the fairness of existing predictive modeling techniques used to support educational practices [19, 15, 48, 55]. Typically, these studies have focused on evaluating the fairness of predictive models that modeled student performance [19, 25, 31]. For example, Gardner et al. [19] proposed the Absolute Between-ROC Area (ABROCA) metric to measure the unfairness of a predictive model as its differential prediction accuracy between different groups of students. Compared to other group fairness metrics (e.g., a demographic parity measure), ABROCA was designed based on equalized odds which ensures equal false and true positive rates among baseline and comparison classes, and therefore avoids individual unfair outcomes in the group fairness measure. By applying ABROCA, Gardner et al. [19] evaluated the unfairness of five mainstream models developed to predict the likelihood of a student to complete a Massive Open Online Course (MOOC). In addition to MOOC education, a group of similar studies has been conducted in other educational settings like higher education [25, 29, 31, 57] and virtual learning environments [41]. In a different vein, Doroudi and Brunskill [15] investigated whether the existing models used for knowledge tracing generate inequitable results for different groups of students and found that the additive factor model was superior to the Bayesian knowledge tracing algorithm and the N-Consecutive Correct Responses heuristic algorithm in delivering fair predictions. Besides, Loukina et al. [32] first discussed different types of fairness that could be applied to evaluate ML models used in educational research, and then utilised both simulated and real datasets to depict how models used for automated scoring of English language proficiency tests might disadvantage students whose first language was not English.

3 Method

3.1 Dataset

The dataset used in this study comprised 3,703 randomly-selected discussion posts created by students in the Learning Management System Moodle at `Monash`

`University`. The topics covered by these posts included arts, design, business, economics, computer science, and mechanical engineering. Here, we differentiated posts as *content-relevant* (e.g., "What is poly-nominal regression?") and *content-irrelevant* (e.g., "When is the due date to submit the second assignment?"). All posts were first manually labeled by a junior teaching staff and then reviewed by two senior teaching staff to ensure the reliability of the derived labels. The dataset contains 2,339 (63%) content-relevant posts and 1,364 (37%) content-irrelevant posts. Additionally, we obtained for each post a student's demographic information, i.e., sex (female or male) and first language (any language). Inspired by [32], which demonstrated that English-as-second-language speakers could be disadvantaged by algorithms used for assessing their learning performance, we transformed the first language categorical variable to a binary form, i.e., English-as-first-language speakers vs. English-as-second-language speakers. The descriptive statistics of the dataset are given in Table 1, based on which we can observe that female students tended to generate more elaborated posts than male students and, similarly, the posts generated by English-as-first-language students were likely to compose more words than those generated by English-as-second-language students.

Table 1: The descriptive statistics of the dataset used in this study. The columns **Male**, **Female**, **First language**, and **Second language** show the number of forum posts generated by students who are male, female, English-as-first-language speakers, and English-as-second-language speakers, respectively.

	All	**Male**	**Female**	**First language**	**Second language**
# Posts	3,703	1,478	2,225	1,585	2,112
# Words	485,737	171,768	308,087	230,806	254,931
# Avg. words / post	131.39	116.77	138.90	145.62	120.71
# Unique words	268,824	97,004	170,171	125,297	143,527
# Avg. unique words / post	72.71	65.94	76.72	79.05	67.96

3.2 Model Selection

As summarized in Sec. 2, both traditional ML models and up-to-date DL models have been exploited to construct automatic classifiers of educational forum posts. Therefore, to enable a comprehensive evaluation, we selected the representative models from both of the two categories in this study.

Traditional ML models. Four traditional ML models were evaluated in this study, namely Naïve Bayes, SVM, Random Forests, and Logistic regression. These models have been widely applied in the context of educational forum classification in previous studies [2, 10, 18, 33, 53]. Relying upon an extensive feature engineering, these models achieved high classification accuracy. To ensure the ML models in our study were comparable with models reported in previous

studies, we replicated the feature engineering process used in previous models, including (i) the top-1000 most frequent unigrams and bigrams contained in the discussion posts [2, 13, 38, 46, 51, 52, 58]; (ii) the length of a post [35, 42, 53, 58]; (iii) the TF-IDF (term frequency-inverse document frequency) score related to each selected unigram [2, 5]; (iv) the frequency of words indicating different psychological processes along with each post (e.g., affects and cognitive process), which were extracted with the aid of LIWC [2, 10, 18, 30, 33, 37, 53]; (v) scores extracted by applying Coh-Metrix to indicate text coherence, linguistic complexity, text readability, and lexical category [30, 37], and (vi) the LSA score indicating the average sentence similarity within a post [30]. In total, 3180 features were engineered as input to the four traditional ML models.

DL models. Existing studies based on DL models typically made use of two types of deep neural networks, i.e., Bi-directional LSTM (Bi-LSTM) [9, 21, 54] and CNN-LSTM [21, 50, 54]. However, it should be noted that the training of these complex neural networks often requires a large amount of annotated data (tens of thousands at least). In recent years, the development of pre-trained language models (e.g., BERT [14]) enabled researchers to exploit the power of these complex neural networks even when there is only a small amount of annotated data available. In more details, a widely-adopted method is to couple a task model (e.g., Bi-LSTM and CNN-LSTM in our case) on top of the output layer of BERT and then use the annotated data to co-train BERT and the task model as a whole to produce the classification results. Given the limited number of annotated posts in our dataset, we also used BERT to empower Bi-LSTM and CNN-LSTM for our classification task.

3.3 Evaluation Metrics

Accuracy metrics. In line with previous studies on constructing automatic classifiers of educational forum posts, we adopted the following four metrics to measure the prediction accuracy of a classifier: Accuracy, Cohen's κ, AUC, and F1 score.

Fairness metrics. To our knowledge, [19] was the first study which attempted to investigate appropriate metrics to evaluate the fairness of predictive models in the field of educational research. Specifically, a metric called Absolute Between-ROC Area (ABROCA) was presented in [19] to measure the prediction unfairness of a predictive model against different demographic groups, which is calculated by finding the definite integral between the ROC curves of the two observed groups. Noticeably, ABROCA has two advantages: 1) ABROCA accounts for performance difference across the entire range of thresholds, which is superior over other fixed-threshold approaches; and 2) ABROCA can be easily computed from prediction results with no need for collecting additional data or computing additional metrics. Therefore, we also used this metric in our study. Notice here, the lower an ABROCA value is, the less algorithmic unfairness a predictive model has.

3.4 Study Setup

Text pre-processing. We pre-process the text contained in a post by performing the following steps: 1) removing invalid characters; 2) removing stopwords; and 3) applying word stemming with the help of the Python package `NLTK` [8].

Model implementation. We used the Python package `scikit-learn` to implement the traditional ML models. To develop DL models, we first generated text embeddings by using the tool `Bert-as-service`[4]. Next, we implemented CNN-LSTM and Bi-LSTM by replicating model parameters reported in previous studies [9, 21, 50, 54]. In CNN-LSTM, we used 128 convolution filters with the width of 5. For both CNN-LSTM and Bi-LSTM, (i) we set the number of hidden units used in the final classification layer to 1 and L2 regularization lambda to 0.001, and utilised sigmoid as the activation function; (ii) the LSTM layer was set to have 128 hidden states and 128 cell states; (iii) we set the batch size to 32 and the maximum input text to 512; (iv) we applied the one cycle policy for training and set the maximum learning rate to 2e-05; (v) the dropout probability was set to 0.5; and (vi) we opted for 50 maximum training epochs with shuffling performed at the end of each epoch together with early stopping mechanism.

Model training. Prior to training a model, we first randomly selected 20% of the available posts as the testing data, and then prepared the training data from the remaining posts. It is worth noting that, as reported in Table 1, the number of posts generated by female and male students are unequal (same for those generated by students with English as first/second language). Previous studies (e.g., [55]) suggested that the algorithmic unfairness of a predictive model may be partially attributed to the unequal amount of training data related to different demographic groups. Therefore, we trained the six classifiers with two different training data samples, namely (i) *original training sample*, i.e., all of the remaining posts (after selecting the testing data) were used as the training data; and (ii) *equal training sample*, i.e., an equal number of posts for each demographic group were randomly selected from the remaining posts and then combined as the training data. It should be pointed out that the same testing data was used to evaluate classification performance in the two training data samples, and thus the results were comparable. While training the models, 10% of the training data was randomly selected as the validation data and the best model was selected based on the error reported in the validation data.

4 Results

Results on original training sample. Table 2 presents the performance of the six classifiers when using original training sample. Based on Table 2, we can have several important observations. When measuring the accuracy of the classification results, DL models were universally superior to traditional ML models, which was in line with the findings presented in previous works [11, 54]. However,

[4] https://github.com/hanxiao/bert-as-service

when scrutinizing the fairness of these models, traditional ML models tended to slightly outperform DL models. For instance, SVM displayed lowest level of unfairness to students of different sex and Logistic Regression achieved the best level of fairness towards students of different first-language backgrounds. These findings suggested that prediction accuracy should not be the only criterion when selecting a predictive model, and more importantly, the fairness of the model should also be evaluated and taken into account. Overall, CNN-LSTM achieved the best prediction accuracy (ranked 1st in both AUC and Cohen's κ) while maintaining acceptable level of fairness to different demographic groups of students (ranked 3rd in both ABROCA (Sex) and ABROCA (Language)). In fact, this implied that a strict accuracy-for-fairness trade-off was not evident in our study. Due to the limited space, the results of Accuracy and F1 score are omitted here, though similar findings can be drawn on those results.

Table 2: Results on original training sample. The top 3 best results are in bold.

Models	**AUC**	**Cohen's κ**	**ABROCA (Sex)**	**ABROCA (Language)**
Random Forests	0.763	0.525	0.038	**0.033**
Naïve Bayes	0.752	0.502	0.062	0.084
Logistic Regression	0.758	0.516	**0.014**	**0.032**
SVM	**0.786**	**0.577**	**0.007**	0.069
CNN-LSTM	**0.795**	**0.584**	**0.014**	**0.063**
Bi-LSTM	**0.786**	**0.565**	**0.010**	0.068

As showed before, there was no indication that DL models produced fairer results than the traditional ML models did. This was not expected given that feature engineering in traditional ML models involved more manual work than the automatic embedding generation in DL models, and therefore might be more susceptible to bias. This indicates that feature engineering may be only marginally related to the unfairness in classification models. However, we also note that in this study we utilised the features extensively engineered in previous studies to address the same classification task, which may have reduced bias. We also observe that the mean ABROCA value for sex (0.024) is only about a half of the mean ABROCA value for language (0.058), which means that the language group (English-as-first-language vs. English-as-second-language speakers) had far more unfair prediction than the sex group. This indicates that linguistic difference of different demographic groups of students may play an important role in improving model fairness. In Figure 1, we also note that, except for Naïve Bayes, all other models provided better classification performance (measured by ROC) to English-as-first-language students. One possible explanation is that these models relied heavily on students' English proficiency to make accurate prediction and therefore posed strong unfairness to students whose first language was not English. Also, this may be partially due to the fact that popular feature and embedding extraction tools were typically trained by using standard English

corpus (e.g., LIWC and BERT). Therefore, it may be worthy allocating further research efforts to scrutinize whether there exist any algorithmic unfairness in these tools and further improve these tools.

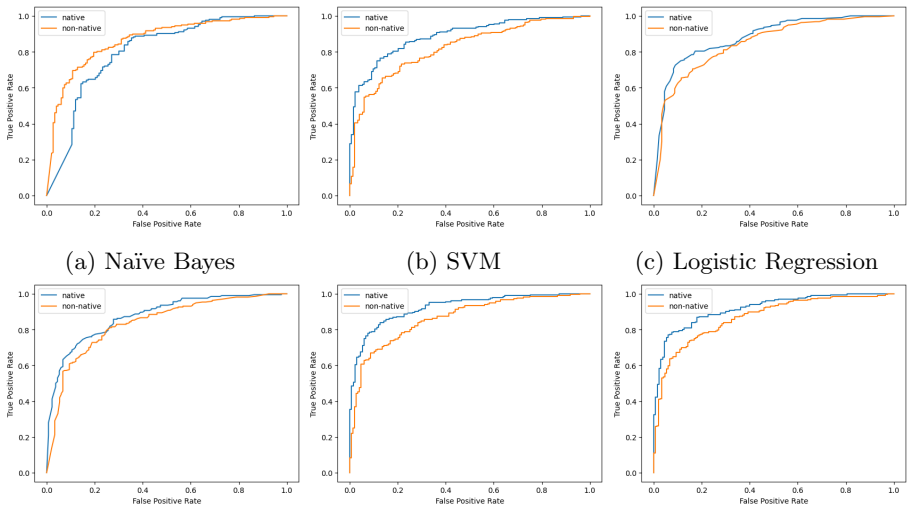
Original training sample vs. equal training sample. In Table 3, we summarized the results of using equal training sample. While the prediction accuracy remained comparable to those of using the original training sample, most of the classifiers (except for Logistic Regression) became fairer in both of the Sex and Language groups after including an equal number of posts related to each demographic group in the training data. In particular, Naïve Bayes had over 61% reduction in ABROCA between male and female, which shows a non-trivial role of data sampling in reducing the algorithmic unfairness of classification models. Therefore, we note that future model training should take demographic balancing into account to encourage fairer classification.

Table 3: Results on equal training sample. The top 3 best results are in bold.

	Sex			**Language**		
Models	**AUC**	**Kappa**	**ABROCA**	**AUC**	**Kappa**	**ABROCA**
Random Forests	0.760	0.518	0.030	0.773	0.545	**0.023**
Naïve Bayes	0.763	0.531	0.024	0.766	0.537	0.052
Logistic Regression	0.783	0.568	**0.003**	0.775	0.547	**0.043**
SVM	0.788	**0.581**	**0.004**	0.772	0.548	**0.012**
CNN-LSTM	**0.792**	**0.579**	0.009	**0.802**	**0.601**	0.062
Bi-LSTM	**0.791**	0.575	**0.007**	**0.784**	**0.559**	0.066

5 Discussion and Conclusion

This paper investigated both the accuracy and fairness of six popular automatic classifiers of educational forum posts. For each classifier, we evaluated the algorithmic fairness between students of different sex and first languages. Our results showed that while classification accuracy varied slightly, the difference of the model unfairness (measured by ABROCA) was more evident. Besides, we observed that, compared to the posts generated by English-as-second-language students, posts generated by English-as-first-language students were overwhelmingly predicted with higher accuracy by most of the classifiers (except for Naïve Bayes). Our results indicated that existing classifiers and feature engineering approaches, originally developed to process standard English, might be prone to discrimination against English-as-second-language students. This finding supported the recent initiatives in the NLP research community to expand the language varieties in the training text corpora to mitigate biases that emerged when researchers designed a prediction model in one context (e.g., texts written by English-as-first-language users) and applied it in another context [7, 45]. As an attempt to address model unfairness, we applied equal sampling to model training. The results were promising with most of the models showing improved fairness. Since it did not require a complex alternation to existing model training,



(a) Naïve Bayes (b) SVM (c) Logistic Regression

(d) Random Forests (e) CNN-LSTM (f) Bi-LSTM

Fig. 1: ROC plots by first-language backgrounds (i.e., English-as-first-language vs. English-as-second-language speakers) on original training sample.

equal sampling can be incorporated with a minimal cost in the future versions of educational forum post classifiers.

Implications. Our findings suggested that a strict performance-for-fairness trade-off is not evident, and by utilising techniques such as equal sampling, researchers can help alleviate the problem of model unfairness without sacrificing classification performance. Moreover, existing model evaluation should take fairness metrics into consideration and avoid models that display a high level of unfairness. We also note that limited work has been done to evaluate bias in feature engineering and embedding extraction in educational research. Future work thus can investigate the possibility of extracting fairer features and evaluating feature fairness before evaluating model fairness, to prevent models from receiving discriminating input, particularly as this information is usually hard to detect later in the model implementation. Additionally, pre-trained language models such as BERT should incorporate a variety of base textual data into their training sets, rather than just using standard English corpus (e.g., Wikipedia).

Limitations. We acknowledged the following limitations of our study. First, the analysis involved only one prediction task, i.e., classifying content-relevant and content-irrelevant forum posts. To further increase the generalizability of our findings, additional prediction tasks using different datasets need to be investigated. Second, the analysis reported in this paper focused only on two types of demographic groups of students. In future studies, we will investigate the algorithmic fairness of different models with respect to other demographic groups, e.g., students of different educational backgrounds and minority students.

Bibliography

[1] Agrawal, A., Venkatraman, J., Leonard, S., Paepcke, A.: Youedu: addressing confusion in mooc discussion forums by recommending instructional video clips (2015)

[2] Almatrafi, O., Johri, A., Rangwala, H.: Needle in a haystack: Identifying learner posts that require urgent response in mooc discussion forums. Computers & Education **118**, 1–9 (2018)

[3] Alrajhi, L., Alharbi, K., Cristea, A.I.: A multidimensional deep learner model of urgent instructor intervention need in mooc forum posts. In: Intelligent Tutoring Systems. pp. 226–236. Springer International Publishing (2020)

[4] Atapattu, T., Falkner, K., Tarmazdi, H.: Topic-wise classification of mooc discussions: A visual analytics approach. International Educational Data Mining Society (2016)

[5] Bakharia, A.: Towards cross-domain mooc forum post classification. In: Learning@Scale. pp. 253–256 (2016)

[6] Barbosa, G., Camelo, R., Cavalcanti, A.P., Miranda, P., Ferreira Mello, R., Kovanović, V., Gašević, D.: Towards automatic cross-language classification of cognitive presence in online discussions. In: LAK. pp. 605–614 (2020)

[7] Bender, E.M., Friedman, B.: Data statements for natural language processing: Toward mitigating system bias and enabling better science. Transactions of the Association for Computational Linguistics **6**, 587–604 (2018)

[8] Bird, S.: Nltk: the natural language toolkit. In: Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions. pp. 69–72 (2006)

[9] Brahman, F., Varghese, N., Bhat, S.: Effective Forum Curation via Multitask Learning p. 8 (2020)

[10] Caines, A., Pastrana, S., Hutchings, A., Buttery, P.J.: Automatically identifying the function and intent of posts in underground forums. Crime Science **7**(1), 19 (2018)

[11] Chen, J., Feng, J., Sun, X., Liu, Y.: Co-training semi-supervised deep learning for sentiment classification of mooc forum posts. Symmetry **12**(1), 8 (2020)

[12] Clavié, B., Gal, K.: Edubert: Pretrained deep language models for learning analytics. arXiv preprint arXiv:1912.00690 (2019)

[13] Cui, Y., Wise, A.F.: Identifying content-related threads in mooc discussion forums. In: Learning@Scale. pp. 299–303 (2015)

[14] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)

[15] Doroudi, S., Brunskill, E.: Fairer but not fair enough on the equitability of knowledge tracing. In: Proceedings of the 9th International Conference on Learning Analytics & Knowledge. p. 335–339. LAK19, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3303772.3303838, `https://doi.org/10.1145/3303772.3303838`

[16] Duran, R., Haaranen, L., Hellas, A.: Gender differences in introductory programming: Comparing moocs and local courses. In: Proceedings of the 51st ACM Technical Symposium on Computer Science Education. pp. 692–698 (2020)
[17] Feng, L., Liu, G., Luo, S., Liu, S.: A transferable framework: Classification and visualization of mooc discussion threads. In: International Conference on Neural Information Processing. pp. 377–384. Springer (2017)
[18] Ferreira, M., Rolim, V., Ferreira Mello, R., Lins, R.D., Chen, G., Gašević, D.: Towards automatic content analysis of social presence in transcripts of online discussions. In: LAK. pp. 141–150 (2020)
[19] Gardner, J., Brooks, C., Baker, R.: Evaluating the fairness of predictive student models through slicing analysis. In: LAK. pp. 225–234 (2019)
[20] Guo, P.J.: Non-native english speakers learning computer programming: Barriers, desires, and design opportunities. In: Proceedings of the 2018 CHI conference on human factors in computing systems. pp. 1–14 (2018)
[21] Guo, S.X., Sun, X., Wang, S.X., Gao, Y., Feng, J.: 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 (2019)
[22] Hew, K.F., Cheung, W.S.: Students' and instructors' use of massive open online courses (moocs): Motivations and challenges. Educational research review **12**, 45–58 (2014)
[23] Hew, K.F., Hu, X., Qiao, C., Tang, Y.: What predicts student satisfaction with moocs: A gradient boosting trees supervised machine learning and sentiment analysis approach. Computers & Education **145**, 103724 (2020)
[24] Hone, K.S., El Said, G.R.: Exploring the factors affecting mooc retention: A survey study. Computers & Education **98**, 157–168 (2016)
[25] Hutt, S., Gardner, M., Duckworth, A.L., D'Mello, S.K.: Evaluating fairness and generalizability in models predicting on-time graduation from college applications. International Educational Data Mining Society (2019)
[26] Jacobs, A.: Two cheers for web u. New York Times **162**(56113), 1–7 (2013)
[27] Jansen, R.S., van Leeuwen, A., Janssen, J., Conijn, R., Kester, L.: Supporting learners' self-regulated learning in massive open online courses. Computers & Education **146**, 103771 (2020)
[28] Khan, A., Ibrahim, I., Uddin, M.I., Zubair, M., Ahmad, S., Firdausi, A., Dzulqarnain, M., Zaindin, M.: Machine learning approach for answer detection in discussion forums: An application of big data analytics. Scientific Programming **2020** (2020)
[29] Kizilcec, R.F., Lee, H.: Algorithmic fairness in education (2020)
[30] Kovanović, V., Joksimović, S., Waters, Z., Gašević, D., Kitto, K., Hatala, M., Siemens, G.: Towards automated content analysis of discussion transcripts: A cognitive presence case. In: LAK. pp. 15–24 (2016)
[31] Lee, H., Kizilcec, R.F.: Evaluation of fairness trade-offs in predicting student success (2020)
[32] Loukina, A., Madnani, N., Zechner, K.: The many dimensions of algorithmic fairness in educational applications. In: Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational

Applications. pp. 1–10. Association for Computational Linguistics, Florence, Italy (Aug 2019). https://doi.org/10.18653/v1/W19-4401, https://www.aclweb.org/anthology/W19-4401

[33] Lui, M., Baldwin, T.: Classifying user forum participants: Separating the gurus from the hacks, and other tales of the internet. In: Proceedings of the Australasian Language Technology Association Workshop 2010. pp. 49–57 (2010)

[34] Mitra, S.K.: Internationalization of education in india: Emerging trends and strategies. Asian Social Science **6**(6), 105 (2010)

[35] Moreno-Marcos, P.M., Alario-Hoyos, C., Muñoz-Merino, P.J., Estévez-Ayres, I., Kloos, C.D.: Sentiment analysis in moocs: A case study. In: 2018 IEEE Global Engineering Education Conference (EDUCON). pp. 1489–1496. IEEE (2018)

[36] Morris, L.V., Finnegan, C., Wu, S.S.: Tracking student behavior, persistence, and achievement in online courses. The Internet and Higher Education **8**(3), 221–231 (2005)

[37] Neto, V., Rolim, V., Ferreira, R., Kovanović, V., Gašević, D., Lins, R.D., Lins, R.: Automated analysis of cognitive presence in online discussions written in portuguese. In: Proceedings of the 13th European Conference on Technology Enhanced Learning. pp. 245–261. Springer (2018)

[38] Ntourmas, A., Avouris, N., Daskalaki, S., Dimitriadis, Y.: 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). pp. 1–8. IEEE (2019)

[39] Probyn, M.: Teachers voices: Teachers reflections on learning and teaching through the medium of english as an additional language in south africa. International Journal of Bilingual Education and Bilingualism **4**(4), 249–266 (2001)

[40] Rayyan, S., Seaton, D.T., Belcher, J., Pritchard, D.E., Chuang, I.: Participation and performance in 8.02 x electricity and magnetism: The first physics mooc from mitx. arXiv preprint arXiv:1310.3173 (2013)

[41] Riazy, S., Simbeck, K., Schreck, V.: Fairness in learning analytics: Student at-risk prediction in virtual learning environments. In: CSEDU (1). pp. 15–25 (2020)

[42] Rossi, L.A., Gnawali, O.: Language independent analysis and classification of discussion threads in coursera mooc forums. In: Proceedings of the 2014 IEEE 15th International Conference on Information Reuse and Integration (IEEE IRI 2014). pp. 654–661. IEEE (2014)

[43] Rovai, A.P.: Sense of community, perceived cognitive learning, and persistence in asynchronous learning networks. The Internet and Higher Education **5**(4), 319–332 (2002)

[44] Rovai, A.P.: Facilitating online discussions effectively. The Internet and Higher Education **10**(1), 77–88 (2007)

[45] Shah, D., Schwartz, H.A., Hovy, D.: Predictive biases in natural language processing models: A conceptual framework and overview. arXiv preprint arXiv:1912.11078 (2019)

[46] Sun, C., Li, S.w., Lin, L.: Thread structure prediction for mooc discussion forum. In: International Conference of Pioneering Computer Scientists, Engineers and Educators. pp. 92–101. Springer (2016)
[47] Toven-Lindsey, B., Rhoads, R.A., Lozano, J.B.: Virtually unlimited classrooms: Pedagogical practices in massive open online courses. The internet and higher education **24**, 1–12 (2015)
[48] Tsai, Y.S., Perrotta, C., Gašević, D.: Empowering learners with personalised learning approaches? agency, equity and transparency in the context of learning analytics. Assessment & Evaluation in Higher Education **45**(4), 554–567 (2020). https://doi.org/10.1080/02602938.2019.1676396, `https://doi.org/10.1080/02602938.2019.1676396`
[49] Verstegen, D., Dailey-Hebert, A., Fonteijn, H., Clarebout, G., Spruijt, A.: How do virtual teams collaborate in online learning tasks in a mooc? International Review of Research in Open and Distributed Learning **19**(4) (2018)
[50] Wei, X., Lin, H., Yang, L., Yu, Y.: A convolution-lstm-based deep neural network for cross-domain mooc forum post classification. Information **8**(3), 92 (2017)
[51] Wise, A.F., Cui, Y., Jin, W., Vytasek, J.: Mining for gold: Identifying content-related mooc discussion threads across domains through linguistic modeling. The Internet and Higher Education **32**, 11–28 (2017)
[52] Wise, A.F., Cui, Y., Vytasek, J.: Bringing order to chaos in mooc discussion forums with content-related thread identification. In: LAK. pp. 188–197 (2016)
[53] Xing, W., Tang, H., Pei, B.: Beyond positive and negative emotions: Looking into the role of achievement emotions in discussion forums of moocs. The Internet and Higher Education **43**, 100690 (2019)
[54] Xu, Y., Lynch, C.F.: What do you want? applying deep learning models to detect question topics in mooc forum posts? In: Wood-stock'18: ACM Symposium on Neural Gaze Detection. pp. 1–6 (2018)
[55] Yan, S., Kao, H.t., Ferrara, E.: Fair class balancing: Enhancing model fairness without observing sensitive attributes. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management. pp. 1715–1724 (2020)
[56] Yang, D., Wen, M., Howley, I., Kraut, R., Rose, C.: Exploring the effect of confusion in discussion forums of massive open online courses. In: Learning@Scale. pp. 121–130 (2015)
[57] Yu, R., Li, Q., Fischer, C., Doroudi, S., Xu, D.: Towards accurate and fair prediction of college success: evaluating different sources of student data. In: EDM. pp. 292–301. ERIC (2020)
[58] Zeng, Z., Chaturvedi, S., Bhat, S.: Learner affect through the looking glass: Characterization and detection of confusion in online courses. International Educational Data Mining Society (2017)