Bigger Data or Fairer Data? Augmenting BERT via Active Sampling for Educational Text Classification

Lele Sha, Yuheng Li, Dragan Gašević and **Guanliang Chen***
Centre for Learning Analytics, Monash University
Melbourne, Victoria, Australia
{lele.sha1, yuheng.li, Dragan.Gasevic, Guanliang.Chen}@monash.edu

Abstract

Pretrained Language Models (PLMs), though popular, have been diagnosed to encode bias against protected groups in the representations they learn, which may harm the prediction fairness of downstream models. Given that such bias is believed to be related to the amount of demographic information carried in the learned representations, this study aimed to quantify the awareness that a PLM (i.e., BERT) has regarding people's protected attributes and augment BERT to improve prediction fairness of downstream models by inhibiting this awareness. Specifically, we developed a method to dynamically sample data to continue the pretraining of BERT and enable it to generate representations carrying minimal demographic information, which can be directly used as input to downstream models for fairer predictions. By experimenting on the task of classifying educational forum posts and measuring fairness between students of different *gender* or *first-language backgrounds*, we showed that, compared to a baseline without any additional pretraining, our method improved not only fairness (with a maximum improvement of 52.33%) but also accuracy (with a maximum improvement of 2.53%). Our method can be generalized to any PLM and demographic attributes. All the codes used in this study can be accessed via https://github.com/lsha49/FairBERT_deploy.

1 Introduction

Pretrained Language Models (PLMs) have been increasingly applied to tackle various NLP tasks in recent years (Li et al., 2019; Yoon et al., 2019; Chan and Fan, 2019; Araci, 2019; Zhang et al., 2019). Along with the wide application of PLMs is growing concerns about the bias encoded in the representations generated by these PLMs (Jin et al., 2020; Lu et al., 2020). For instance, de Vassimon Manela et al. (2021) demonstrated that stereotypical associations were encoded in PLMs when tackling a pronoun resolution task. More importantly, such bias has been demonstrated to be harmful to the prediction *fairness* of downstream models, i.e., there exists a consistent gap between the prediction accuracy for people of different protected attributes. For example, Minot et al. (2021) showed that the gender-related bias embedded in PLMs could be propagated to downstream classification tasks in medical scenarios.

As a remedy, researchers have endeavored to develop techniques to *debias* PLMs. These techniques, more often than not, focused on the fine-tuning stage when using a PLM, e.g., correcting the bias hidden behind the learned representations by removing associations between embedding features and protected attributes during the fine-tuning process or using a protected-attribute-balanced dataset to fine-tune a PLM (de Vassimon Manela et al., 2021). It is worth noting that, though assuming that the bias contained in a PLM is associated with the amount of demographic information carried in the learned representations, these debiasing techniques oftentimes failed to (i) explicitly quantify the capability of a PLM in revealing people's demographic attributes or (ii) depict the relationship between the amount of demographic information contained in the learned representations and the prediction fairness in downstream tasks.

Inspired by the studies which demonstrated the benefits of using additional task-specific data to continue to pretrain a language model and boost prediction accuracy (Araci, 2019; Clavié and Gal, 2019; Chalkidis et al., 2020; Shen et al., 2021; Beltagy et al., 2019), in this study, we focused on continuing the pretraining of a language model (i.e., BERT (Devlin et al., 2018)) with carefully-selected data so as to reduce the amount of demographic information contained in the learned representations and subsequently enhance the downstream predic-

* Corresponding author.

Proceedings of the 29th International Conference on Computational Linguistic, pages 1275–1285
October 12–17, 2022.

tion models in terms of both *accuracy* and *fairness*. Our rationale is essentially in line with those held in (de Vassimon Manela et al., 2021; Minot et al., 2021), i.e., the bias carried in a PLM can be potentially reduced by inhibiting a PLM's awareness of people's protected attributes. Formally, this study was guided by the following **R**esearch **Q**uestions:

RQ1 To what extent can the representations generated by a vanilla BERT[1] predict people's protected attributes?

RQ2 To what extent can BERT's awareness of protected attributes be inhibited by actively sample data to continue its pretraining?

RQ3 What are the impacts of inhibiting BERT's awareness of protected attributes on the prediction fairness in the downstream model?

We based our study on the task of classifying discussion forum posts in education, which is widely recognized as important in assisting instructors to provide timely support to students, especially in courses with a high student-teacher ratio (Ntourmas et al., 2021; Wei et al., 2017). This study was approved by the Human Research Ethics Committee at Monash university (Project ID 30074). We used a dataset consisting of over 228K forum posts generated by students when undertaking their studies at the same university, and information about students' *gender* and *first-language backgrounds* were also contained in the dataset. To answer RQ1, we used the representations generated by the vanilla BERT as input to a logistic regression model to predict students' protected attributes. To answer RQ2, building upon studies on Active Learning (AL), we proposed a data sampling method to selectively sample data, which contain minimal information about protected attributes, to continue to pretrain BERT, after which we measured whether the learned representations became less capable of revealing students' protected attributes. To answer RQ3, after applying additional pretraining to BERT, we further used the the learned representations as input to a different logistic regression model to predict the categorical label of a forum post (i.e., content relevant or irrelevant). Through extensive evaluations, we demonstrated that the proposed sampling method can effectively identify data to decrease a PLM's awareness of protected attributes and enhance predictive models used in downstream tasks in terms of both accuracy and fairness. Our sampling method can be generalized to any PLM or protected attributes.

[1] A vanilla BERT refers to one without any additional pretraining.

2 Related Work

2.1 Bias in Pretrained Language Models

PLMs have been documented to contain biases against certain socio-demographic groups (e.g., black and female), which was partially caused by the use of low-quality data when constructing a PLM (Lucy and Bamman, 2021; Nadeem et al., 2020; de Vassimon Manela et al., 2021). Nadeem et al. (2020) showed that harmful stereotypes commonly existed in online text. When using such online texts for training, PLMs can easily pick up harmful stereotypes and act against the disadvantaged groups. For instance, when predicting the emotion polarity and toxicity of a piece of text, PLMs are prone to classify text written by females as more emotional than those written by their male counterpart (Jin et al., 2020; Touileb et al., 2021; Silva et al., 2021; Bhardwaj et al., 2021; Mozafari et al., 2020). Another reason is that the data generated by disadvantaged groups was less used in constructing a PLM and thus causing these disadvantaged groups to be under-represented compared to other groups. When detecting hate speech, texts written in African American English dialect were more likely to be mistakenly classified than texts written in standard English (Halevy et al., 2021).

2.2 Debiasing Pretrained Language Models

Existing studies in this strand of research mostly stressed on the fine-tuning stage when using a PLM (Bhardwaj et al., 2021; Silva et al., 2021; Jin et al., 2020). Among these studies, a majority of them aimed to debias the learned representations by regularizing the BERT model during fine-tuning. For instance, Bhardwaj et al. (2021) proposed a method to identify and remove the semantic features which contained sensitive information (e.g., gender-related) when propagating through BERT layers, thereby reducing BERT-induced bias in the downstream tasks. Silva et al. (2021) applied a loss regularizer where a loss is incorporated in training to minimize bias learned during fine-tuning. Alternatively, some researchers attempted to debias a PLM by modifying the task-specific data samples before using them to fine-tune the PLM (Prost et al., 2019; Islam et al., 2021; Pruksachatkun et al.,

2021). Typically, the data was modified with the aim of removing traits that are indicative of people's gender (de Vassimon Manela et al., 2021; Minot et al., 2021), race (Mozafari et al., 2020), or dialect (Mozafari et al., 2020). These data modification methods have been demonstrated effective in enhancing the prediction fairness of downstream models in pronoun resolution (de Vassimon Manela et al., 2021) and toxicity detection (Mozafari et al., 2020). In addition, researchers demonstrated that a PLM could be debiased by enabling the downstream task model to work with other models in an ensemble-based manner (Halevy et al., 2021). For example, Halevy et al. (2021) showed that by adding a specialized classifier trained by text written in the African American English dialect to the ensemble framework, the model displayed fewer racial biases when detecting toxic language compared to using PLM alone.

Our work shared a similar rationale with (Gajane and Pechenizkiy, 2017; Minot et al., 2021; Mozafari et al., 2020), i.e., by reducing a PLM's awareness of the protected attributes related to data, the PLM is less likely to propagate bias to the model used in a downstream task and the prediction fairness of the task model can be enhanced. However, our work distinguished itself from two aspects. Firstly, instead of focusing on the fine-tuning stage, we focused on debiasing a PLM by actively sampling protected-attribute-uninformative data to continue to pretrain the PLM. Secondly, we explicitly quantified the capability of the PLM in predicting protected attributes and measure its impact on the prediction fairness of the downstream model.

2.3 Pretrained Language Models in Education

Driven by the great success in the broader NLP communities, PLMs have been also applied in solving various tasks in the field of education, such as generating questions for assessment (Lu et al., 2021), providing timely feedback to support student learning (Lin et al., 2022), and scoring answers or essays authored by students (Ormerod et al., 2021). Among these tasks, the classification of forum posts has received lots of attention from researchers due to its important role in facilitating instructors to support students in the era of online learning (Clavié and Gal, 2019; Alrajhi et al., 2020; Geller et al., 2021; Capuano et al., 2021). For instance, Clavié and Gal (2019) further pretrained BERT using forum posts collected in the education domain and classified students' posts to a task label of whether these posts requires urgent attention from instructors or not. The constructed classifiers could assist instructor to quickly identify students that require urgent help. However, these studies did not attempt to quantify or alleviate the impact from the bias hidden behind PLMs. Considering that education is often regarded as a high-stake commodity, we were thus motivated to investigate the bias of PLM based on the task of classifying student-generated forum posts.

3 Method

3.1 Dataset and Models

Dataset. The dataset used in this study was retrieved from the Learning Management System at an Australian university. The original dataset consisted of 291,242 student-generated posts in discussion forums when undertaking courses of Information Technology, Engineering, Education, Business and Economics, etc. In addition, we obtained students' demographic information including their gender (female vs. male) and first language, which enabled us to investigate prediction fairness from gender and first-language backgrounds perspectives. Inspired by (Loukina et al., 2019), which demonstrated that English-as-second-language students could be disadvantaged by algorithms used for assessing their learning performance, we categorised students according to their *first-language backgrounds* as either English-as-first-language or English-as-second-language students. After filtering posts containing less than 5 words, there were 228,903 posts left, from which we randomly selected 3,703 posts and manually annotated them as either *content-relevant* (e.g., "*What is poly-nominal regression?*") and *content-irrelevant* (e.g., "*When is the due date to submit the assignment?*"). Each post was first labeled by a junior teaching staff and then reviewed by two senior teaching staff to ensure the reliability of the derived labels. There are 2,339 (63%) content-relevant posts and 1,364 (37%) content-irrelevant posts. We denoted these 3,703 posts as **Annotated Data**. Recall that part of our goal was to reveal the capability of BERT in predicting students' demographic attributes, we therefore randomly select 5% of the remaining 225,200 unannotated posts and used them as an indepedent data set (denoted as **Demographic Data**) to scrutinize how BERT would differ in predicting demographic attributes after undergoing additional

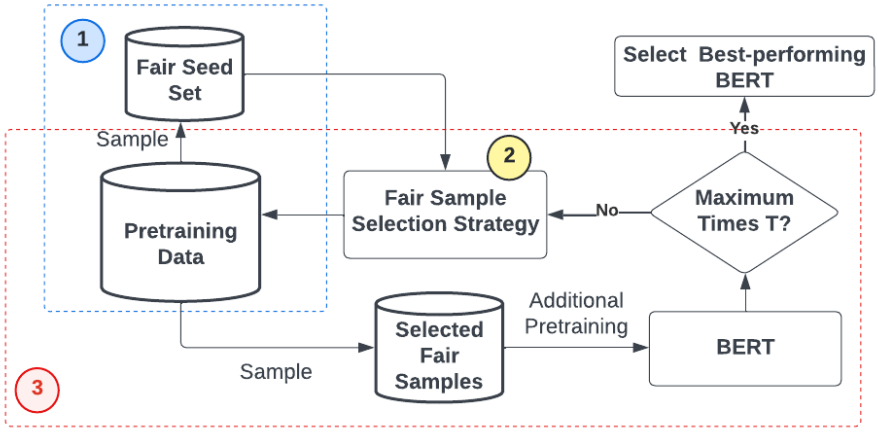


Figure 1: The fair active sampling method proposed in this study.

Table 1: Dataset statistics. The columns **Male**, **Female**, **First**, and **Second** show the number of posts made by students who are male, female, English-as-first-language, and English-as-second-language students, respectively.

		All	Male	Female	First	Second
Pretraining	# Posts	213,940	84,564	140,374	97,255	127,945
	# Words	32,264,332	12,212,745	20,015,047	13,674,012	18,590,320
	# Avg. words / post	143.27	144.42	142.58	140.60	145.30
Demographic	# Posts	11,260	4,509	6,751	4,955	6,305
	# Words	1,522,127	568,810	953,308	712,231	809,877
	# Avg. words / post	135.18	126.15	141.21	143.74	128.45
Annotated	# 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

pretraining. Lastly, the remaining 213,940 unannotated posts were used as the data pool from which we sampled instances to apply additional pretraining to BERT (denoted as **Pretraining Data**). The statistics of the three sets are given in Table 1.

Models. A variety of effective models have been used to classify a forum post in education (Bakharia, 2016; Capuano et al., 2021; Almatrafi et al., 2018). Given that our main goal was to investigate whether the changes in the BERT-generated representations in terms of the amount of demographic information would produce any impact on the prediction fairness in a downstream task, we adopted logistic regression to take the BERT-generated post embeddings as input to distinguish the different labels of forum posts for simplicity. We denoted this model as `LR-Task`. Similarly, we also used logistic regression to predict the demographic attributes of the creator of a post, i.e., one logistic regression model for each demographic attribute. We denoted these two models as `LR-Demographic`.

3.2 Fair Active Sampling

Our data sampling method was partially inspired by the studies on Active Learning (AL) (Anahideh et al., 2022; Castro and Nowak, 2008). AL is a well-investigated strategy used to train a supervised machine learning model by enabling the model to proactively request to identify and label new high-value data samples to facilitate its training process and achieve better prediction accuracy. Considering the strong capacity of AL in enhancing various machine learning models, we were interested whether it could be used to alleviate the bias contained in a PLM by selecting fair data (i.e., those containing little information about students' protected attributes) to continue to pretrain the PLM. There are three key steps in our method, as depicted in Figure 1 and detailed below.

Step 1: Fair Seed Set Construction (i.e., ① depicted in Figure 1). When applying AL, researchers often first randomly select a small subset from the unlabeled data pool and labeled them to initialize the training of a machine learning model. Then,

this subset of labeled data can be referenced as a seed set to sample more informative data (e.g., those would help the model to reduce the maximum training error) to continue the model training. In our case, as our goal was to enhance a PLM's fairness by reducing the demographic information contained in the learned representations, we aimed to construct a *fair* seed set from the annotated data pool to guide the subsequent active data sampling. To this end, we expected the fair seed set to contain zero *distribution bias* and minimal *hardness bias* (Yan et al., 2020; Smith et al., 2014). Distribution bias refers to the unequal distribution of different student groups (e.g., female vs. male) within each prediction task label (e.g., content relevant vs. content irrelevant). Formally, it can be defined as the the difference of probabilities of prediction task label $Y = y$, conditioned upon the value of a protected attribute G where $G \in \{0, 1\}$ and $Y \in \{0, 1\}$:

$$D(y) = Pr(Y = y \mid G = 1) - Pr(Y = y \mid G = 0) \quad (1)$$

A seed set that contains zero distribution bias indicates $D(y) = 0$. On the other hands, Hardness bias measures to what extent data instances contained in a dataset are difficult to be correctly labeled. That is, if a data instance does not share the same task label with most of its *k-nearest neighbors*, then it tends to be difficult to correctly label this instance. Similar to (Smith et al., 2014; Yan et al., 2020), we used the metric *k-Disagreeing Neighbors* (kDN) to measure the local overlap of a data instance x with its k-nearest neighbors (identified by calculating the Euclidean distance between their vanilla BERT-generated representations) regarding their task labels. We chose $k = 5$ to calculate the kDN of an instance (as suggested in (Yan et al., 2020)). A large kDN (close to 1) indicates that the data instance is difficult to be correctly classified. If the kDN distribution of a student group (e.g., female) is larger than that of the other group (e.g., male), there exists Hardness bias (denoted as H) between the two groups, which can be calculated by applying the Jensen-Shannon (JS) distance, as defined below:

$$H(y) = JS(\{f(x, y) \mid G = 1\} - \{f(x, y) \mid G = 0\}) \quad (2)$$

where $G \in \{0, 1\}$, $Y \in \{0, 1\}$, and $f(x, y)$ is the kDN distribution of data instances with $G = g$ and $Y = y$. To keep the Hardness bias between two protected groups minimal, we constructed the fair seed set by adding the instances in an one-by-one manner, and a data instance could only be included if the overall Hardness bias of the fair seed set was decreasing after including the instance.

Step 2: Query Strategy Selection (i.e., ② depicted in Figure 1). With the seed set constructed, we further enriched it with more fair instances by calculating the instances' informativeness with respect to students' protected attributes. Specifically, we adopt three representative query strategies in the AL research to enrich the seed set, which all support the selection of multiple instances at a time and have been proven to be effective in assisting machine learning models to achieve better prediction performance. All the machine learning models used by these strategies were built based on the seed set constructed above and took the vanilla BERT-learned representation of a post as input to predict the protected attribute of the post creator.

- `Query By Committee (QBC)` (Dagan and Engelson, 1995), which first trained an ensemble of models, i.e., ten logistic regression classifiers in our case. Each of the classifiers was built based on a random subset of the fair seed set. If the ensemble of models could not reach an agreement on the predicted protected attribute, then the post was selected.

- `Learning Active Learning (LAL)` (Konyushkova et al., 2017), which trained a random forest regressor to predict the expected error reduction (derived based on the predicted probabilities of different protected attributes) for all unlabeled samples in Pretraining Data, and select samples with the most error reduction.

- `Least Confident Classification (LCC)` (Settles and Craven, 2008; Bilgic et al., 2010; Tong and Chang, 2001), which trained a classifier based on logistic regression to predict the protected attribute of a post creator and selected samples with the least confidence, i.e., the predicted probabilities should be as close to 0.5 as possible in the binary protected attribute classification problems in our study.

Step 3: Dynamic and Fair Sampling (i.e., ③ depicted in Figure 1). Every time when a query strategy described above is applied, $K\%$ fraction of the Pretraining Data will be sampled and used

to perform additional pretraining to BERT. Instead of setting a large value to K and only perform one-time sampling, which may hinder the identification of effective samples to augment a PLM, we designed a dynamic approach to sample instances for multiple times (i.e., with small values for K) and, more importantly, the number of fair instances specific to a protected group can be determined based on the current prediction bias existed in the downstream task. Let t denote the number of times that BERT has undergone additional pretraining, G_0 and G_1 denote the two protected groups, and the current prediction accuracy in the downstream task (i.e., those with the aid of the latest BERT) for G_0 and G_1 are denoted as acc_0 and acc_1, respectively. Then, for the $(t+1)$-th pretraining, the ratio between the number of sampled instances between G_0 and G_1 is $acc_1 : acc_0$, i.e., the group with a lower prediction accuracy will have more samples. The above sampling process is repeated until it reaches the maximum times (denoted as T_{max}) allowed to performed additional pretraining and then the best-performing BERT in terms of prediction fairness is selected.

3.3 Study Setup

Baselines. In addition to the three variants of the proposed sampling method (denoted as `AL-QBC`, `AL-LAL`, and `AL-LCC`, respectively), we implemented three baselines for comparisons: (i) `w/o Pretraining`, in which LR-Task takes as input the vanilla BERT-learned representations (i.e., without any additional pretraining); (ii) `Random`, in which LR-Task takes as input the learned representations from BERT with randomly-selected samples for additional pretraining; and (iii) `Equal-demographic`, in which LR-Task takes as input the learned representations from BERT with additional pretraining based on data an equal representation of different protected groups (but not considering their representations in terms of different task labels).

Model implementation and training. We used the BERT model provided by *huggingface*[2]. The AL methods were implemented by *alipy*[3] and the logistic regression models by *sklearn*[4]. For the fair data sampling method we proposed, we set the size of the fair seed set (Step 1 of our proposed approach) to be 500, the value of T_{max} to be 6 (i.e., the number of times that BERT could have additional pretraining), and the value of K to be 5% (i.e., the fraction of data to be sampled from the Pretraining Data by an AL strategy every time). To guide the data sampling process of our proposed method (i.e., Step 3), we constructed an independent *reference subset* by using the same method described in Step 1 to sample 500 posts from the Annotated Data, and this subset was used to measure the changing prediction bias existed between different protected groups throughout the whole debiasing process. When training the logistic regression model for predicting students' protected attributes (i.e., LR-Demographic), we randomly split the Demographic Data into training and testing sets in the ratio of 8:2 and the hyper-parameters were determined via 5-fold cross-validation. We used a similar method to construct the logistic regression model for predicting the types of posts (i.e., LR-Task), but based on the remaining 3,203 labeled posts in Annotated Data (after constructing the reference subset described above). Our procedure to continue the pretraining of BERT is in line with similar studies (Devlin et al., 2018; Araci, 2019; Beltagy et al., 2019), with the maximum length of input text as 512, learning rate as 2e-05, batch size as 16, the maximum number of training epochs as 12, and early stopping mechanisms were used.

Evaluation metrics. The capability of BERT in discerning students' protected attributes can be indirectly measured in terms of the prediction accuracy of LR-Demographic, which is measured by using AUC. Recall that we aimed to reduce the demographic information carried in BERT-learned representations; therefore, a lower AUC of LR-demographic indicates better results (i.e., a lower awareness of protected attributes). We also used the AUC to measure the prediction accuracy of LR-Task in distinguishing different types of forum posts, but a higher AUC is preferable for LR-Task. To measure the prediction bias held by LR-Task, we adopted the Absolute Between-ROC Area (ABROCA) metric proposed by (Gardner et al., 2019), which has been widely used to measure algorithmic bias in relevant educational studies (Riazy et al., 2020; Tsai et al., 2020; Paquette et al., 2020). A lower ABROCA indicates better prediction fairness.

[2]https://huggingface.co/
[3]http://parnec.nuaa.edu.cn/
[4]https://scikit-learn.org/

Table 2: Results of LR-Demographic and LR-Task with the aid of different AL strategies. Here, T_{max} denotes the maximum number of times that BERT received additional pretraining. The number inside brackets indicates improvement compared to the results of `w/o pretraining`. LR-demo is short for LR-Demographic. The best result in each metric is in bold. The signs ↑ and ↓ are used to indicate whether a higher (↑) or lower (↓) value is more preferred in a metric.

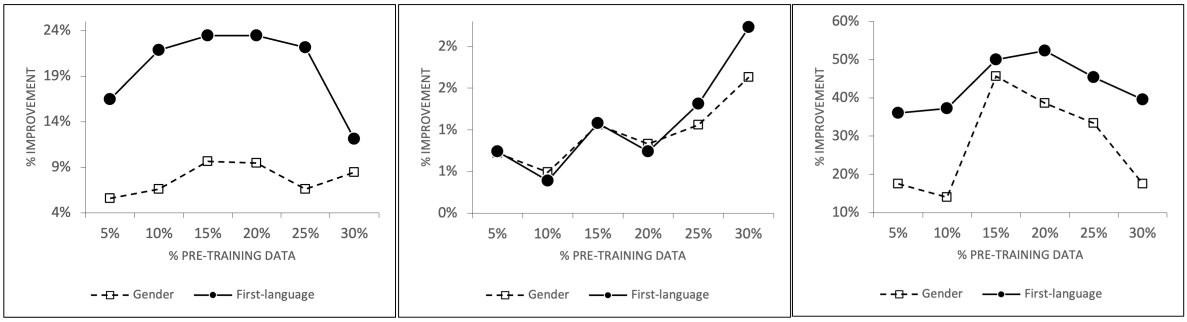
Row ID	Methods	T_{max}	First-language backgrounds			Gender		
			LR-Demo	LR-Task		LR-Demo	LR-Task	
			↓ AUC	↑ AUC	↓ ABROCA	↓ AUC	↑ AUC	↓ ABROCA
1	w/o pretraining	-	0.686	0.869	0.086	0.591	0.882	0.057
2	Ramdom	1	0.692 (-0.87%)	0.876 (0.81%)	0.098 (-13.95%)	0.611 (-3.38%)	0.892 (1.13%)	0.089 (-56.14%)
3	Equal		0.670 (2.33%)	0.883 (1.66%)	0.079 (8.14%)	0.595 (-0.68%)	0.889 (0.84%)	0.066 (-15.79%)
4	AL-QBC		0.591 (13.85%)	0.879 (1.15%)	0.105 (-22.09%)	0.559 (5.41%)	0.889 (0.77%)	0.059 (-3.51%)
5	AL-LAL		0.589 (14.14%)	0.876 (0.85%)	0.069 (19.77%)	0.552 (6.60%)	0.898 (1.85%)	0.055 (3.51%)
6	AL-LCC		0.573 (16.47%)	0.878 (1.01%)	0.055 (36.05%)	0.558 (5.58%)	0.891 (1.02%)	0.047 (17.54%)
7	Ramdom	6	0.688 (-0.29%)	0.889 (2.30%)	0.112 (-30.23%)	0.588 (0.51%)	0.895 (1.47%)	0.072 (-26.32%)
8	Equal		0.621 (9.48%)	0.889 (2.30%)	0.095 (-10.47%)	0.561 (5.08%)	0.889 (0.84%)	0.066 (-15.79%)
9	AL-LCC		**0.525 (23.47%)**	**0.891 (2.53%)**	**0.041 (52.33%)**	**0.534 (9.64%)**	**0.899 (1.96%)**	**0.031 (45.61%)**

4 Results

Results on RQ1. The results of LR-demographic are given in Table 2 (Row 1). For both of the two protected attributes, LR-Demographic achieved an AUC score larger than 0.5, which implies that the representations of forum posts learned by BERT did carry informative features that can be utilized by machine learning models to reveal the demographic attributes of the students. In particular, the AUC value of first-language backgrounds (0.686) is much higher than that of gender (0.591). A possible explanation is that, students in the groups of different first-language backgrounds, compared to the groups of a different gender, tend to display more linguistic differences in their forum posts, which can be captured by BERT and used to reveal their demographic attributes.

Results on RQ2. To answer RQ2, we implemented the three AL strategies described in Section 3.2. To compare their effectiveness in inhibiting BERT's awareness of protected attributes, we first set T_{max} as 1, i.e., only performing one-time additional pretraining, and the results are given in Table 2 (Row 2-6, Column LR-Demo). Based on the results, we can observe that, by simply using an equal number of instances generated by different protected groups for additional pretraining (i.e., the baseline `Equal`), it cannot guarantee that BERT's awareness of protected attributes can be reduced. However, all of the AL strategies proposed in this work showed effectiveness in removing demographic information embedded in the vanilla BERT-learned representations, especially when dealing with the first-language groups (with an improvement of 13.85%~16.47%), where vanilla BERT demonstrated a higher awareness of students' demographic attributes. Notice that, for both of the two protected attributes, the best results were delivered by `AL-LCC`. Then, we re-ran evaluation of `AL-LCC` and increased the value of T_{max} to 6, and also compared the results to those of `Random` and `Equal` with the same T_{max} value (Row 7-9 in Table 2). Here, the AUC of LR-Demo was even reduced to very close to 0.5 (i.e., 0.525) in the first-language groups and 0.534 in the gender groups, which provides strong evidence of the effectiveness of our fair sampling method in stopping BERT to record demographic attributes into its learned representations.

Results on RQ3. To measure the impact of BERT's reduced awareness of protected attributes on the prediction fairness in the downstream task, we measured the performance of LR-Task which took as input the representations learned by the vanilla BERT or BERT with additional pretraining. The results are also given in Table 2, based on which we can have observations similar to those made when analyzing the AUC of LR-Demographic. Firstly, in the results of `Random` and `Equal` (Row 2-3 and Row 7-8), which randomly sampled data or only maintained a balance in terms of protected attributes in the selected data, the prediction bias tended to be exacerbated and the prediction fairness could be worsen up to 30.23%. Secondly, among the three AL strategies, `AL-LCC` outperformed the other two in enhancing prediction fairness, even when there was only one-time additional pretraining (i.e., $T_{max} = 1$, Row 6 in Table 2), and the



(a) LR-Demo (AUC) (b) LR-Task (AUC) (c) LR-Task (ABROCA)

Figure 2: The relative improvements of AL−LCC compared to w/o Pretraining from $T_{max} = 1$ to $T_{max} = 6$.

ABROCA improvement was up to 36.05% for first-language groups and 17.54% for gender groups. When T_{max} was increased to 6 (Row 9 in Table 2), the improvement was further boosted to 52.33% and 45.61% for the two groupings. Thirdly, in addition to the increased prediction fairness, we observe that the prediction accuracy was boosted with slight improvements of 2.53% and 1.96% for the two groupings (Row 9 in Table 2), respectively, compared to those of Random (2.30% and 1.47%). This further demonstrates that, by carefully selecting the data used to perform additional pretraining, both the prediction accuracy and fairness can be simultaneously enhanced.

To investigate the amount of data needed to reach the maximum prediction fairness, we plotted the relative improvement achieved by AL−LCC compared to the baseline w/o Pretraining across the whole additional pretraining process (Figure 2). We found that, with the aid of AL−LCC, LR-Demographic achieved the best AUC when using 15%~20% of the available unlabeled data (i.e., $T_{max} = 3$ or $T_{max} = 4$). This implies that, instead of using all the available data, it can be more effective and efficient by carefully selecting a subset of them for additional pretraining.

5 Discussion and Conclusion

To debias BERT by directly reducing the information related to protected attributes in the learned representations, this study developed a dynamic and fair sampling method to select data to perform additional pretraining to BERT, which is capable of significantly inhibiting BERT's awareness of protected attributes and subsequently improved both the prediction fairness and accuracy in the downstream task. Here, we further elaborated on our study's practical insights and implications and discussed the limitations to be addressed in the future.

Implications. Firstly and most importantly, our study corroborated the findings of previous studies (de Vassimon Manela et al., 2021; Minot et al., 2021), i.e., prediction fairness in a downstream task can be greatly enhanced by reducing a PLM's awareness of sensitive protected attributes, e.g., the amount of information related to such protected attributes in the learned representations. Second, as demonstrated in Table 2, only two out of the three AL strategies used in this study could enhance the prediction fairness in downstream tasks, suggesting that it is of extreme importance to select the appropriate measure for sample informativeness in terms of protected attributes. Third, our study demonstrated that, by carefully selecting fair samples to further pretrain a PLM, even only with 15%~20% of the available unlabeled data, not only the prediction fairness but also the prediction accuracy can be enhanced. This implies that prediction accuracy can benefit from keeping prediction fairness as part of the goal when performing additional pretraining to a PLM.

Limitations. Firstly, the effectiveness of the proposed fair data sampling method was only validated based on BERT and one dataset in the field of education. Future studies are needed to replicate the study using other PLMs or datasets to further validate the presented findings. Secondly, we focused on debiasing BERT in terms of two protected attributes (i.e., first-language backgrounds and gender) separately. Future work may further investigate methods to debias a PLM by considering other types of protected attributes or simultaneously taking multiple of them into consideration. Lastly, we only experimented with a limited set of values for the parameters used in the proposed method, e.g., $K = 5\%$ (the fraction of available unlabeled data

to be sampled for additional pretraining). In the future, it would be worthwhile to develop methods to automatically determine the best values for such parameters.

References

Omaima Almatrafi, Aditya Johri, and Huzefa Rangwala. 2018. Needle in a haystack: Identifying learner posts that require urgent response in mooc discussion forums. *Computers & Education*, 118:1–9.

Laila Alrajhi, Khulood Alharbi, and Alexandra I Cristea. 2020. A multidimensional deep learner model of urgent instructor intervention need in mooc forum posts. In *International conference on intelligent tutoring systems*, pages 226–236. Springer.

Hadis Anahideh, Abolfazl Asudeh, and Saravanan Thirumuruganathan. 2022. Fair active learning. *Expert Systems with Applications*, 199:116981.

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Aneesha Bakharia. 2016. Towards cross-domain mooc forum post classification. In *Proceedings of the Third (2016) ACM Conference on Learning@ Scale*, pages 253–256.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Rishabh Bhardwaj, Navonil Majumder, and Soujanya Poria. 2021. Investigating gender bias in bert. *Cognitive Computation*, 13(4):1008–1018.

Mustafa Bilgic, Lilyana Mihalkova, and Lise Getoor. 2010. Active learning for networked data. In *ICML*.

Nicola Capuano, Santi Caballé, Jordi Conesa, and Antonio Greco. 2021. Attention-based hierarchical recurrent neural networks for mooc forum posts analysis. *Journal of Ambient Intelligence and Humanized Computing*, 12(11):9977–9989.

Rui Castro and Robert Nowak. 2008. Active learning and sampling. In *Foundations and Applications of Sensor Management*, pages 177–200. Springer.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*.

Ying-Hong Chan and Yao-Chung Fan. 2019. A recurrent bert-based model for question generation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 154–162.

Benjamin Clavié and Kobi Gal. 2019. Edubert: Pretrained deep language models for learning analytics. *arXiv preprint arXiv:1912.00690*.

Ido Dagan and Sean P Engelson. 1995. Committee-based sampling for training probabilistic classifiers. In *Machine Learning Proceedings 1995*, pages 150–157. Elsevier.

Daniel de Vassimon Manela, David Errington, Thomas Fisher, Boris van Breugel, and Pasquale Minervini. 2021. Stereotype and skew: Quantifying gender bias in pre-trained and fine-tuned language models. In *EACL 2021-16th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*, pages 2232–2242. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Pratik Gajane and Mykola Pechenizkiy. 2017. On formalizing fairness in prediction with machine learning. *arXiv preprint arXiv:1710.03184*.

Josh Gardner, Christopher Brooks, and Ryan Baker. 2019. Evaluating the fairness of predictive student models through slicing analysis. In *Proceedings of the 9th international conference on learning analytics & knowledge*, pages 225–234.

Shay A Geller, Nicholas Hoernle, Kobi Gal, Avi Segal, Hyunsoo Gloria Kim, David Karger, Marc Facciotti, Kamali Sripathi, and Michele Igo. 2021. New methods for confusion detection in course forums: Student, teacher and machine. *IEEE Transactions on Learning Technologies*.

Matan Halevy, Camille Harris, Amy Bruckman, Diyi Yang, and Ayanna Howard. 2021. Mitigating racial biases in toxic language detection with an equity-based ensemble framework. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–11.

Rashidul Islam, Kamrun Naher Keya, Ziqian Zeng, Shimei Pan, and James Foulds. 2021. Debiasing career recommendations with neural fair collaborative filtering. In *Proceedings of the Web Conference 2021*, pages 3779–3790.

Xisen Jin, Francesco Barbieri, Brendan Kennedy, Aida Mostafazadeh Davani, Leonardo Neves, and Xiang Ren. 2020. On transferability of bias mitigation effects in language model fine-tuning. *arXiv preprint arXiv:2010.12864*.

Ksenia Konyushkova, Raphael Sznitman, and Pascal Fua. 2017. Learning active learning from data. *Advances in neural information processing systems*, 30.

Liangyou Li, Xin Jiang, and Qun Liu. 2019. Pretrained language models for document-level neural machine translation. *arXiv preprint arXiv:1911.03110*.

Jionghao Lin, Mladen Rakovic, David Lang, Dragan Gasevic, and Guanliang Chen. 2022. Exploring the

politeness of instructional strategies from human-human online tutoring dialogues. In *LAK22: 12th International Learning Analytics and Knowledge Conference*, pages 282–293.

Anastassia Loukina, Nitin Madnani, and Klaus Zechner. 2019. The many dimensions of algorithmic fairness in educational applications. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–10, Florence, Italy. Association for Computational Linguistics.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2020. Gender bias in neural natural language processing. In *Logic, Language, and Security*, pages 189–202. Springer.

Owen HT Lu, Anna YQ Huang, Danny CL Tsai, and Stephen JH Yang. 2021. Expert-authored and machine-generated short-answer questions for assessing students learning performance. *Educational Technology & Society*, 24(3):159–173.

Li Lucy and David Bamman. 2021. Gender and representation bias in gpt-3 generated stories. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 48–55.

Joshua R Minot, Nicholas Cheney, Marc Maier, Danne C Elbers, Christopher M Danforth, and Peter Sheridan Dodds. 2021. Interpretable bias mitigation for textual data: Reducing gender bias in patient notes while maintaining classification performance. *arXiv preprint arXiv:2103.05841*.

Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2020. Hate speech detection and racial bias mitigation in social media based on bert model. *PloS one*, 15(8):e0237861.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*.

Anastasios Ntourmas, Sophia Daskalaki, Yannis Dimitriadis, and Nikolaos Avouris. 2021. Classifying mooc forum posts using corpora semantic similarities: a study on transferability across different courses. *Neural Computing and Applications*, pages 1–15.

Christopher M Ormerod, Akanksha Malhotra, and Amir Jafari. 2021. Automated essay scoring using efficient transformer-based language models. *arXiv preprint arXiv:2102.13136*.

Luc Paquette, Jaclyn Ocumpaugh, Ziyue Li, Alexandra Andres, and Ryan Baker. 2020. Who's learning? using demographics in edm research. *Journal of Educational Data Mining*, 12(3):1–30.

Flavien Prost, Nithum Thain, and Tolga Bolukbasi. 2019. Debiasing embeddings for reduced gender bias in text classification. *arXiv preprint arXiv:1908.02810*.

Yada Pruksachatkun, Satyapriya Krishna, Jwala Dhamala, Rahul Gupta, and Kai-Wei Chang. 2021. Does robustness improve fairness? approaching fairness with word substitution robustness methods for text classification. *arXiv preprint arXiv:2106.10826*.

Shirin Riazy, Katharina Simbeck, and Vanessa Schreck. 2020. Fairness in learning analytics: Student at-risk prediction in virtual learning environments. In *CSEDU (1)*, pages 15–25.

Burr Settles and Mark Craven. 2008. An analysis of active learning strategies for sequence labeling tasks. In *proceedings of the 2008 conference on empirical methods in natural language processing*, pages 1070–1079.

Jia Tracy Shen, Michiharu Yamashita, Ethan Prihar, Neil Heffernan, Xintao Wu, Sean McGrew, and Dongwon Lee. 2021. Classifying math knowledge components via task-adaptive pre-trained bert. In *International Conference on Artificial Intelligence in Education*, pages 408–419. Springer.

Andrew Silva, Pradyumna Tambwekar, and Matthew Gombolay. 2021. Towards a comprehensive understanding and accurate evaluation of societal biases in pre-trained transformers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2383–2389.

Michael R Smith, Tony Martinez, and Christophe Giraud-Carrier. 2014. An instance level analysis of data complexity. *Machine learning*, 95(2):225–256.

Simon Tong and Edward Chang. 2001. Support vector machine active learning for image retrieval. In *Proceedings of the ninth ACM international conference on Multimedia*, pages 107–118.

Samia Touileb, Lilja Øvrelid, and Erik Velldal. 2021. Using gender-and polarity-informed models to investigate bias. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 66–74.

Yi-Shan Tsai, Carlo Perrotta, and Dragan Gašević. 2020. 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.

Xiaocong Wei, Hongfei Lin, Liang Yang, and Yuhai Yu. 2017. A convolution-lstm-based deep neural network for cross-domain mooc forum post classification. *Information*, 8(3):92.

Shen Yan, Hsien-te Kao, and Emilio Ferrara. 2020. Fair class balancing: Enhancing model fairness without observing sensitive attributes. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 1715–1724.

Wonjin Yoon, Jinhyuk Lee, Donghyeon Kim, Minbyul Jeong, and Jaewoo Kang. 2019. Pre-trained language model for biomedical question answering. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 727–740. Springer.

Haoyu Zhang, Jianjun Xu, and Ji Wang. 2019. Pretraining-based natural language generation for text summarization. *arXiv preprint arXiv:1902.09243*.