JOURNAL OF LEARNING ANALYTICS

SOLAR SOCIETY for LEARNING ANALYTICS RESEARCH

Volume 13(1), 143–162. https://doi.org/10.18608/jla.2026.8905

Learning Analytics to Uncover Ethnic Bias in Educational Texts: An Ensemble Learning Approach

Josmario Albuquerque[1], Bart Rienties[2], Martin Hlosta[3], Wayne Holmes[4]

Abstract

Online learning platforms have expanded access to education but also raise concerns about biased content, particularly in text-based learning materials such as textbooks, lesson plans, and course excerpts. Such biases can perpetuate discrimination, can harm student outcomes, and can often be difficult to detect, as identification typically relies on time-consuming human review. Learning analytics (LA) can enhance this process by supporting human reviewers through automated detection, offering a scalable solution while retaining human judgment for nuanced evaluations. Accordingly, this LA study explores two research questions: *RQ1: Which features might support the identification of ethnic bias in text-based online learning materials?* and *RQ2: Which classification approaches might be suitable for identifying ethnic bias in text-based online learning materials?* First, we identified features signalling potential ethnic bias (presence or absence) in textual content using a dataset ($N = 345$) labelled by 193 students from diverse ethnic backgrounds. Then, we evaluated multiple machine learning (ML) models for their effectiveness in bias classification. The results suggest significant correlations between perceived bias and content from social sciences. Additionally, through bootstrap analysis, support vector machines and random forest classifiers showed consistent performance in bias identification (with F1-scores of 0.71 and 0.70 on the test set, respectively). In contrast, the naive Bayes (NB) model demonstrated the highest precision (0.75 on the test set). We discuss these findings and their implications for LA, emphasizing the importance of quality and inclusive educational tools. As an initial step toward automated bias classification in education, this study provides a foundation for spotting ethnic bias in learning content, supporting fairer technologies for more inclusive learning environments.

Notes for Research and Practice

- There are statistically significant correlations between ethnic bias and social sciences content.
- Random forest (RF) and stacking (STK) classifier models were more reliable for ethnic bias classification.
- The naive Bayes (NB) model is recommended in scenarios that prioritize precision in bias detection.
- An extended labelled dataset is made available for promoting fairer artificial intelligence (AI) applications.
- A baseline approach for more inclusive learning analytics (LA) in online environments is provided.

Keywords

Learning analytics, ethnic bias, machine learning, online learning, open educational resources.

Submitted: 14/02/2025 — **Accepted:** 23/11/2025 — **Published:** 25/02/2026

[1] *Corresponding author E-mail: josmario.albuquerque@gmail.com Address: Institute of Educational Technology, The Open University, Walton Hall, Kents Hill, Milton Keynes, MK7 6AA, UK. ORCID iD: https://orcid.org/0000-0002-7437-0747*

[2] *Corresponding author E-mail: bart.rienties@open.ac.uk Address: Institute of Educational Technology, The Open University, Walton Hall, Kents Hill, Milton Keynes, MK7 6AA, UK. ORCID iD: https://orcid.org/0000-0003-3749-9629*

[3] *E-mail: martin.hlosta@ffhs.ch Address: Institute for Distance Learning and eLearning Research, Swiss Distance University of Applied Sciences, Brig, SwitzerlandKnowledge Media Institute, The Open University, Milton Keynes, UK. ORCID iD: https://orcid.org/0000-0002-7053-7052*

[4] *E-mail: wayne.holmes@ucl.ac.uk Address: Institute of Education, University College London, London, UK. ORCID iD: https://orcid.org/0000-0002-8352-1594*

1. Introduction

Online learning platforms have expanded learning opportunities to diverse learners internationally. This expansion can be attributed to the availability of online educational content, particularly text-based materials such as textbooks, lesson plans, and video transcriptions (Mayer, 2019). However, these resources often reflect systemic biases inherent in educational systems, affecting students from various socioeconomic backgrounds and demographic groups (Sabnis et al., 2022; Pennington et al.,

2016). By "bias" we mean intergroup bias—an inclination in favour of (positive bias) or against (negative bias) individuals from certain groups based on their social attributes, such as gender (gender bias) or ethnicity (ethnic bias) (Greenwald & Krieger, 2006).

These biases can manifest both explicitly—when individuals are aware of them—and implicitly—when they occur automatically, without individuals' awareness (Daumeyer et al., 2019). Both forms can be incorporated into curricula and learning technologies, even without the intention of content creators, which can affect student academic performance and personal development (Skopec et al., 2021; Shahjahan et al., 2022). However, the vast amount of textual data in online platforms, along with cultural and context-dependent aspects of bias, makes the task of identifying these biases complex and demanding for stakeholders, including educators, developers, and policymakers (Allen & Seaman, 2007; Dutt et al., 2017; Pryzant et al., 2020).

While prior research has explored automated bias detection in texts, it primarily focused on non-educational domains like news articles or social media using advanced language models (Pryzant et al., 2020; Mehrabi et al., 2021). These approaches often prioritize classification performance at the expense of transparency, limiting their practical value for human-in-the-loop educational contexts (Holstein & Doroudi, 2019; Baker & Hawn, 2022). In the educational domain, much of the existing work has focused on algorithmic bias in predictive models of student outcomes (Holstein & Doroudi, 2022; Kizilcec & Lee, 2022; Baker & Hawn, 2022), such as course completion rates or academic performance. Techniques such as fairness constraints and transparency tools have been proposed (Karumbaiah & Brooks, 2021; Sloan-Lynch & Morse, 2024), but comparatively little attention has been paid to automating bias flagging in text-based learning materials. Thus, there remains a gap in interpretable, context-aware machine learning (ML) methods designed specifically to automate the detection of bias within learning materials—particularly in ways that reflect the culturally diverse perspectives of student populations.

To bridge this gap, we explored how learning analytics (LA)—"the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimising learning and the environments in which it occurs" (Long et al., 2011, p. 3)—could support the classification of ethnic bias in text-based learning materials. In this study, we position LA as a mechanism not only for understanding learner behaviours but also for uncovering structural inequities in instructional content, enabling fairer learning designs.

The use of LA is motivated by its known potential, evidenced by previous research on complex and data-driven educational issues, such as predicting student performance, comprehending student emotions, and uncovering performance gaps (Namoun & Alshanqiti, 2020; Sedrakyan et al., 2020). Extending this potential, we apply LA principles to the classification of bias within learning materials, focusing on the role of analytics in supporting content review and educational equity initiatives. Furthermore, while we acknowledge the existence of various types of bias in online learning, we decided to focus on ethnic bias because it is a widespread issue shown to impact students in different ways, such as access to opportunities, resources, and social relationships (Ashong & Commander, 2012; Q. Nguyen et al., 2020; Richardson et al., 2020; Tate & Warschauer, 2022).

In summary, this study addresses the following research questions:

- RQ1: Which features might support the identification of ethnic bias in text-based online learning materials?
- RQ2: Which classification approaches might be suitable for identifying ethnic bias in text-based online learning materials?

By focusing on interpretable, feature-driven ML models, we contribute toward extending LA beyond algorithmic bias in predictive modelling—offering new insights to support critical content creation that accounts for multiple student perspectives.

To answer these questions, we adopted a comprehensive methodological approach covering several steps, from the selection of a dataset to the training and evaluation of ML classifiers. Initially, we pre-processed a dataset of open educational resources (OERs) comprising 345 excerpts, previously assessed by higher education students from diverse ethnic backgrounds for potential bias. This dataset was curated from publicly available OER repositories (see Section 3.1). Accordingly, the pre-processing involved feature expansion and selection to identify relevant attributes for bias classification (Section 3.3). Afterwards, we selected a range of ML classifiers, known for their efficacy in similar tasks (Section 3.4). These algorithms underwent a thorough training and fine-tuning phase using a portion of our dataset, followed by a validation step in which they were tested against an independent portion of the data, allowing us to evaluate their ability to generalize and classify ethnic bias in unknown content. More details about this process are provided in Section 3.

2. Literature Review

The following review first explores how intergroup bias manifests in learning environments (Section 2.1). We then examine the role of LA in detecting and addressing bias (Section 2.2), followed by an overview of ML techniques for bias detection (Section 2.3). We also highlight key research gaps and position this study within the existing literature.

2.1 Intergroup Bias and Learning

The growth of online learning platforms has expanded access to learning resources globally. This rise, already noted in earlier studies (Baumeister & Vohs, 2007; Ferguson, 2012), was further accelerated by the COVID-19 pandemic, which required a large-scale shift from face-to-face classes to online learning (Beaunoyer et al., 2020; Bansak & Starr, 2021). To illustrate, Bansak and Starr (2021) reported that in the US alone, this transition impacted approximately 200,000 households. As online education became more prevalent, its challenges became more evident, potentially affecting a larger number of students (Hollister et al., 2022). One significant concern is the increased risk of biases inadvertently being incorporated into learning platforms (Copur-Gencturk et al., 2022). As online platforms gain more users, the likelihood of such biases influencing educators and learners increases, which is a crucial aspect to monitor and address (Baumeister & Vohs, 2007).

Intergroup bias—the tendency to favour individuals from one's own group while holding negative attitudes toward those from different groups (Greenwald & Krieger, 2006; Tajfel, 1970)—can manifest in various ways, including biases based on ethnicity, gender, or religious beliefs. In educational settings, such biases can create disparities in student experiences and outcomes. For instance, Q. Nguyen and colleagues (2020) explored inequalities in distance learning using LA and found that individuals from ethnic minority backgrounds (e.g., African students) were more likely to struggle to complete their online courses. This highlights that biases in online learning can disadvantage certain groups by impacting their learning success.

Biases can also emerge in the design of learning materials. For example, Dawkins and colleagues (2017) analyzed physics questions from online courses and found that some questions exhibited gender bias. Specifically, those involving the interpretation of certain diagrams (e.g., two-dimensional graphs) tended to favour male students. This effect was linked to the way questions were framed, often reflecting contexts where men are overrepresented.

Furthermore, intergroup biases are not always apparent and can be found in the subtle aspects of the learning environment, such as algorithms (Baker & Hawn, 2022; Balica, 2018) and textbooks (Skopec et al., 2021; Mohamed et al., 2020). Baker and Hawn (2022) reviewed the literature on algorithmic bias in education and reported various studies suggesting the presence of those biases in different aspects of learning platforms, such as in predictive models trying to forecast student dropouts and failing rates, in automated essay scoring, and in assessments of language proficiency. In another systematic literature review, Skopec and colleagues (2021) identified over 200 articles on curriculum decolonization—a movement to remove colonial biases and perspectives from educational content. They showed that when those biases are present in the curriculum (e.g., textbooks), such materials can benefit the views of privileged groups (e.g., white individuals) to the detriment of the perspectives of stigmatized populations (e.g., Black individuals).

When those biases are not addressed, they can create an environment of unequal opportunities and discrimination, preventing learners from different backgrounds from achieving their full potential. For example, research shows that the presence of those biases in education can lead to mental health issues, such as anxiety, depression, and negative thinking, impairing learners' performance and motivation (Albuquerque et al., 2017; Jordano & Touron, 2017; Doubé & Lang, 2012). A comprehensive literature review featuring 45 experiments across 38 articles (Pennington et al., 2016) highlighted that biases against various groups, such as women and African-Americans, activate those psychological processes (e.g., anxiety and negative thinking), which in turn adversely impact student academic performance.

Despite the known evidence on the consequences of intergroup biases in online learning, identifying these biases is complex and challenging. First, bias can be subjective and shaped by cultural and moral judgments, as suggested by studies indicating that cultural factors significantly influence perceptions of individuals and biases between groups (Brewer & Yuki, 2007; Chiu et al., 1997; Dovidio & Gaertner, 2010). Second, the vast amount of data in learning environments would require a huge effort to assess potential biases by hand (Dutt et al., 2017). In summary, while humans can identify bias based on insight, this method is not scalable and is highly influenced by individual perspectives.

As we move forward, it is important to explore LA and computational tools to systematically identify intergroup biases. Below, we describe how LA has been increasingly used to support researchers in understanding such complex issues in online learning, improving student engagement, and mitigating performance disparities among different demographic groups.

2.2 LA

Over the past decade, LA has been used to understand student learning behaviours, predict performance, and identify inequalities (F. Chen & Cui, 2020; Namoun & Alshanqiti, 2020; Q. Nguyen et al., 2020; Sedrakyan et al., 2020). Early work predominantly focused on uncovering inequitable patterns. For instance, Q. Nguyen and colleagues (2020) explored how students from ethnic minority backgrounds, particularly Black and African students, faced a higher risk of course non-completion. Similarly, Sabnis and colleagues (2022) analyzed procrastination patterns and found that racial minorities and first-generation students exhibited higher levels of procrastination. These studies highlight how LA can identify patterns of inequity in online learning environments, but merely identifying these patterns is insufficient to ensure fairness or equity.

Ongoing debates have focused on the complexities involved in addressing bias and fairness within predictive analytics, particularly around the use of student demographic data. Holstein and Doroudi (2022) emphasize that predictive models in education often perpetuate existing biases unless explicit steps are taken to mitigate them. There has been substantial debate

regarding whether including protected attributes (such as gender or ethnicity) in predictive models genuinely enhances fairness or inadvertently reinforces stereotypes and discriminatory practices (Holstein & Doroudi, 2022; Kizilcec & Lee, 2022; Yu et al., 2021). For example, Al-Zawqari and Vandersteen (2023) investigated the fairness of predictive models used in online STEM courses, revealing that ignoring demographic information when designing models can actually reduce fairness. Their findings also suggest that explicitly incorporating student-protected attributes (e.g., age, gender) can improve model fairness without significantly sacrificing accuracy. However, researchers have highlighted that the ethical implications and the risk of reinforcing stereotypes or stigmatizing certain groups remain significant concerns (Baker & Hawn, 2022; A. Nguyen et al., 2023).

Additionally, several researchers advocate for comprehensive approaches that move beyond merely adjusting predictive models. To illustrate, Karumbaiah and Brooks (2021) stressed the importance of transparency in algorithmic decision-making to foster trust and ensure fairer educational outcomes. Hutt and colleagues (2022) further highlight that addressing algorithmic bias requires involving educators in interpreting LA outputs and making decisions about interventions to ensure culturally responsive and equitable educational practices. Sloan-Lynch and Morse (2024) exemplify such a participatory approach through tools like the Course Diversity Dashboard, enabling educators to visually explore relationships between demographic characteristics, student behaviours, and academic outcomes, thus supporting targeted interventions for marginalized populations.

Despite these advancements, the existing LA literature predominantly addresses fairness related to predictive models and interventions, with limited attention given to bias inherent in learning materials themselves. Given the substantial influence of educational texts and resources on student perceptions, motivation, and engagement, integrating LA techniques with bias detection methods specific to learning content presents a crucial and under-explored opportunity. The following section reviews how ML techniques have been utilized for bias detection in broader contexts and discusses associated challenges when extending these methods to educational texts.

2.3 ML and Related Work

ML has already been adopted for detecting bias in other domains, including education, hiring, and online content moderation (Mehrabi et al., 2021; Baker & Hawn, 2022). These approaches typically rely on supervised learning models trained on annotated datasets to classify biased versus non-biased content. Recent advances in natural language processing (NLP) have introduced more sophisticated models, including ensemble learning and deep learning, to improve classification accuracy (Pryzant et al., 2020; Albuquerque et al., 2025).

Feature-based methods, which use interpretable linguistic and contextual attributes, offer particular advantages for fairness-centred applications like LA. They enable greater transparency and actionable insights for educational stakeholders, complementing the traditional goals of LA (Holstein & Doroudi, 2019; Baker & Hawn, 2022). Accordingly, this study focuses on exploring feature-driven ML models as an initial step toward systematic, interpretable bias detection in learning materials.

Traditional ML models such as logistic regression (LG), support vector machines (SVMs), and random forests (RFs) seem promising in detecting bias due to their interpretability and relatively low computational cost (F. Chen & Cui, 2020; Sagi & Rokach, 2018). For instance, SVMs have demonstrated strong performance in binary classification tasks with imbalanced datasets, making them potential candidates for identifying subtle biases in learning materials (Dutt et al., 2017). Similarly, RF models have been preferred for their robustness against overfitting and ability to capture complex decision boundaries.

Recent studies have explored NLP-based methods for bias detection. For instance, Pryzant and colleagues (2020) developed an approach to detect and neutralize subjective biases in sentences using ML. While promising, such models primarily focus on word-level biases and do not account for broader contextual cues within paragraphs or documents. Additionally, their models were designed for general applications such as Wikipedia and online news, limiting their applicability to educational settings. Related research has investigated earlier versions of large language models (LLMs) for bias classification in transcriptions of online courses (Albuquerque et al., 2025). While LLMs demonstrated potential for flagging biased content, aligning their outputs with the nuanced, culturally grounded perceptions of diverse student groups remained challenging. Although recent advances such as prompt engineering and retrieval-augmented generation (RAG) techniques offer opportunities to better incorporate context, explainability and bias mitigation in LLMs remain active areas of research (Bommasani et al., 2021; Dai et al., 2024; Kumar et al., 2024).

This study addresses these gaps by adopting a feature-driven ML approach that uses linguistic, psycholinguistic, and contextual markers to classify ethnic bias. By comparing traditional classifiers (SVM, RF, naive Bayes (NB)) with ensemble and stacking (STK) approaches, we evaluate their effectiveness in educational contexts. This approach integrates context awareness and multi-perspective annotations to promote fairer, more interpretable bias detection in learning materials.

3. Methods

Our methodology is guided by principles of human-centred artificial intelligence (AI) and explainable AI, which emphasize creating AI systems that are transparent, interpretable, and aligned with human values (Liao & Varshney, 2021). These paradigms align with our broader aim of developing ethnic bias classification mechanisms that account for the perspectives

of diverse ethnic populations. To this end, we focused on using feature-based ML models that offer greater interpretability compared to deep learning approaches. Although some ensemble methods (e.g., RF) are more complex internally, they still allow practical explanation strategies, such as analyzing feature importance, which can support transparency in content review workflows.

In particular, this study is organized into five parts: (i) An existing dataset of OERs (see Section 3.1), previously assessed by higher education students for potential ethnic biases, was pre-processed to format and expand its features. Afterwards, (ii) features were selected using a filter method based on statistical measures, specifically a correlation-based approach. Then, (iii) various ML classifiers were implemented for ethnic bias classification. These classifiers were then (iv) trained, evaluated, and fine-tuned using the selected features and a portion of the dataset (the "training set"). Finally, (v) the models were tested on an independent portion of the dataset (the "test set") to assess their generalization performance on unseen data.

We also incorporated an ensemble learning strategy by stacking various combinations of classifiers. Ensemble methods are known for combining multiple models to capitalize on their complementary strengths (Sagi & Rokach, 2018), potentially improving performance while maintaining interpretability. By testing these combinations, we aimed to identify the most suitable model or ensemble for ethnic bias classification. This process is illustrated in Figure 1 and detailed in the next sections.

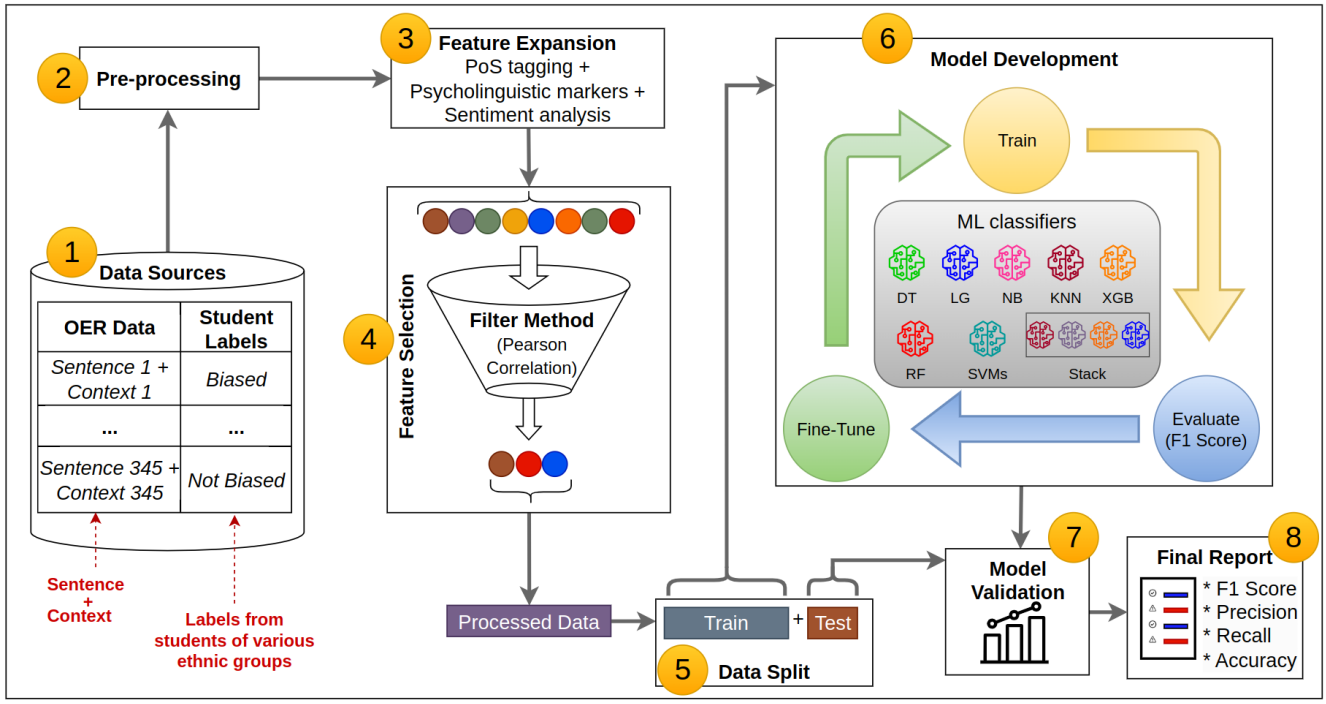

Figure 1. Methods overview illustrating the steps from data extraction to the final bias classifications. This includes data pre-processing, feature expansion and selection, and model development and validation.

3.1 Dataset

The data used in this study were collected in a previous mixed-methods study focused on assessing the perspectives of different ethnic groups about ethnic bias in OERs (Albuquerque, 2023). This particular dataset[1] was selected because it captures diverse perspectives on bias within real educational materials, using a structured annotation protocol that was refined through multiple cycles of design-based research. While full details are available in the aforementioned study, below we provide an overview of the participants, annotation process, key attributes, examples, and pre-processing steps.

Participants and Annotations. The data comprise 345 contextualized sentences extracted from various OERs and annotated by 193 higher education students from the UK and US. Participants self-identified as Asian, Black, mixed, or white. Each participant was presented with a sentence, highlighted within its broader context, and asked to assess potential ethnic bias considering a target ethnic group reference. Figure 2 shows the instructions and illustrates the task.

Key Attributes. The dataset comprises the original sentences, contextual attributes, and the annotations made by students from underrepresented (Group 1) and majority (Group 2) ethnic groups. Specifically, the dataset includes the following attributes:

[1]The dataset is available at https://figshare.com/s/267058706722b161ace5.

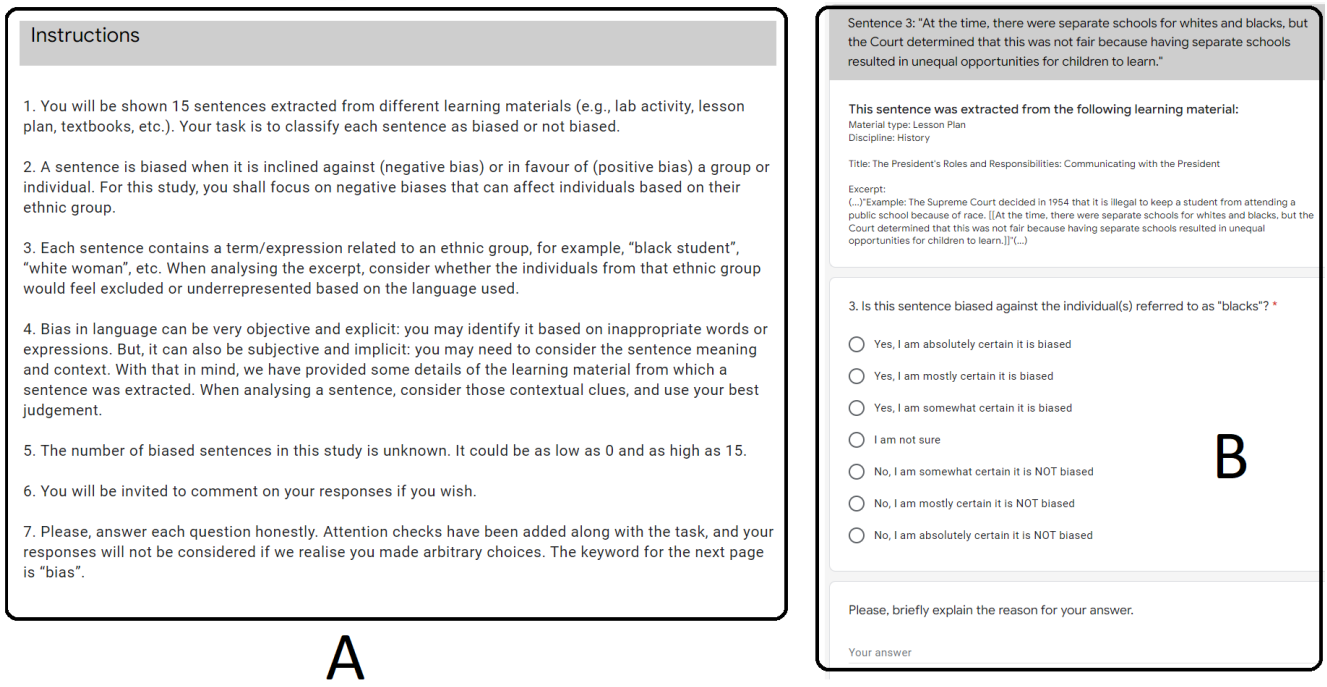

Figure 2. Illustration of the bias labelling task presented to participants. (A) Instructions provided before the task. (B) Example of a sentence shown to participants, with corresponding bias assessment question.

- **oer_title**: the title of the OER from which the sentence was extracted;
- **oer_type**: the type or format of the OER (e.g., lesson plan, course, activity);
- **subject**: the academic discipline or topic associated with the OER;
- **group_reference**: a keyword in the sentence referring to a specific ethnic group, used to assess bias;
- **sentence**: the original sentence extracted from the OER;
- **excerpt**: the broader context surrounding the sentence, where available;
- **cleaned_sent**: a cleaned version of the sentence post pre-processing;
- **Annotations**: counts of "biased," "not biased," and "undecided" marks provided by participants.

Examples of Perceived Bias. To illustrate the nature of perceived biases, some examples from the dataset are included below. While each example was perceived as "biased" by at least the majority of one group, perceptions varied across individuals, reflecting the subjective nature of bias (see Section 2.1). For each example, we briefly summarize a reason participants cited for perceiving it as biased, noting that other interpretations may also exist.

- *"The three housing developments and many of the government loan programs were almost exclusively for whites. Levittown contracts contained covenants restricting occupancy by members other than the Caucasian race."* (Perceived as biased due to its terminology, "Caucasian race.")
- *"For white families, suicide and divorce became more acceptable, while the opposite occurred for Black families."* (Perceived as biased due to generalizations suggesting that all Black families were less accepting of issues like suicide and divorce.)
- *"The franchise or right to vote was being extended to more white males as income-related eligibility requirements were being dropped by more states."* (Perceived as biased because it frames the expansion of voting rights as progress toward white males while overlooking the continued exclusion of women and ethnic minority groups.)

Pre-processing. Two transformations were applied to the original annotations to support feature selection and machine learning. First, a continuous **bias score** was computed as the difference between the proportion of "biased" and "not biased" responses. Scores range from -1 (strong consensus of no bias) to $+1$ (strong consensus of bias), with values near 0 indicating greater disagreement or uncertainty among participants. Second, for binary classification, a final **class label** was created: a sentence was labelled as *biased* if at least one group (Group 1 or Group 2) had a positive bias score (i.e., greater than 0); otherwise, it was labelled as *not biased*. This provided a consolidated label reflecting the perceptions of both majority and underrepresented groups.

In addition to these initial transformations, further feature engineering steps were conducted and detailed in Section 3.3.

3.2 Materials and Tools

Key materials and tools used in this study included the software development environment, libraries, and hardware. For instance, Python 3 was used as the primary programming language given the availability of various libraries supporting ML and data processing. Accordingly, Pandas (The Pandas development team, 2020) was used for data pre-processing, and Scikit-learn (Pedregosa et al., 2011) was adopted for implementing and fine-tuning each model. In addition, the source code was versioned using GitHub (https://github.com/josmarios/ethnic-bias-classification), which benefits reproducibility. In terms of hardware, all analyses were conducted on a standard laptop (CPU: 4 × 2.3-GHz Intel Core i5-6200U; RAM: 16GB).

3.3 Features

To identify potential features for classifying ethnic bias in text-based online learning materials, we examined various linguistic and contextual aspects informed by existing literature. These features were grouped into five categories: psycholinguistic markers, linguistic abstraction, language valence, ethnic group mentions, and contextual attributes. Below, we outline the rationale and relevance of each category.

Psycholinguistic Markers (*psychological*) . Psycholinguistic markers are known to reflect the author's psychological state during writing (Sboev et al., 2015), which can help identify emotive texts that may reveal potential biases. This assumption aligns with Tajfel's intergroup bias framework, emphasizing the emotional significance of social groups (Tajfel, 1970; Tajfel et al., 1971, 2000). Accordingly, we adopted the psycholinguistic features and naming conventions proposed by Sboev and colleagues (2015), including aggressiveness, socialization readiness, emotional stability, and self-reference ratios. More details are provided in Table 1.

Table 1. Psycholinguistic features and respective formulas.

Category	Feature	Formula
psychological	Aggressiveness	$aggressiv = \frac{Number\ of\ verbs}{Number\ of\ words}$
psychological	Emotional Stability	$emo\_stability = \frac{Number\ of\ adjectives + Number\ of\ adverbs}{Number\ of\ nouns + Number\ of\ verbs}$
psychological	Self-Reference Ratio	$self\_ref = \frac{Number\ of\ 1^{st}\text{-}person\ pronouns}{all\ pronouns}$
psychological	Socialization	$socialization = \frac{Number\ of\ verbs}{Number\ of\ nouns}$

Linguistic Abstraction (*abstraction*). Linguistic abstraction refers to the level of specificity or generality in descriptions, often influenced by biases in perceptions of in-group and out-group members (Maass, 1999). For example, abstract terms (e.g., adjectives) may be used differently than concrete terms (e.g., action verbs) to describe individuals. Metrics such as adjective-to-verb ratio, sentence length, and unique word ratio help capture this phenomenon (Tincher et al., 2016). Table 2 lists all features from this category used in this study. (PoS stands for "parts of speech" in the table.)

Table 2. Features related to linguistic abstraction and their respective formulas.

Category	Feature	Formula
abstraction	Adjective Ratio	$adj\_ratio = \frac{Number\ of\ adjectives}{Number\ of\ PoS}$
abstraction	Adverb Ratio	$adv\_ratio = \frac{Number\ of\ adverbs}{Number\ of\ PoS}$
abstraction	Noun Ratio	$noun\_ratio = \frac{Number\ of\ nouns}{Number\ of\ PoS}$
abstraction	Pronoun Ratio	$pnoun\_ratio = \frac{Number\ of\ pronouns}{Number\ of\ PoS}$
abstraction	Sentence Length	$sent\_len = Number\ of\ words$
abstraction	Unique PoS ratio	$unique\_pos = \frac{Number\ of\ unique\ PoS}{Number\ of\ PoS}$
abstraction	Unique Word Ratio	$unique\_word = \frac{Number\ of\ unique\ words}{sent\_len}$
abstraction	Verb Ratio	$vrb\_ratio = \frac{Total\ verbs}{Number\ of\ PoS}$

Language Valence (*valence*). Language valence assesses the emotional tone (positive vs. negative) of words and phrases, which can also reflect biases and stereotypes (Osgood et al., 1957; Mendelsohn et al., 2020). For example, texts describing certain ethnic groups with consistently negative emotional language may signal potential biases. Metrics in this category quantify the positivity, neutrality, or negativity of language in a sentence (see Table 3).

Ethnic Group Mentions (*mentions*). Mentions of specific ethnic groups may correlate with how bias is perceived, depending on the broader context in which they appear. However, group references alone do not necessarily indicate bias and may be more informative when considered alongside other attributes, such as valence (e.g., mentioning one group more frequently and positively than others). This assumption draws on theories of group categorization and intergroup dynamics, which highlight that language referencing social groups can sometimes reinforce stereotypes or biases (Tajfel et al., 2000). To operationalize this potential signal, we used a non-exhaustive keyword list developed in a prior study (Albuquerque, 2023), comprising over 1,000 terms. This list was built through a synonym-based expansion across four ethnic categories (Asian, Black, mixed, and white) and combined with person- and location-related terms (e.g., "teacher," "Africa"). It was used during data collection to pre-screen OERs for ethnic-related content. The full keyword list is available in the code repository linked in Section 3.2.

The features derived from ethnic group mentions are summarized in Table 3.

Table 3. Features for group mentions and valence. All features, except for *target_ethnic* (exclusive to sentences), were computed for both sentences and excerpts.

Category	Feature	Description
mentions	Ethnic Majority Mentions	$ethnic\_maj$ = *Number of mentions of white-related ethnic groups*
mentions	Ethnic Minority Mentions	$ethnic\_min$ = *Number of mentions of ethnic minority groups*
mentions	Target Group	$target\_ethnic$ = *0 (ethnic minority) or 1 (ethnic majority)*
valence	Text Valence	*valence = Vader valence between* -1 *and* $+1$

Contextual Attributes (*context*). Contextual attributes include characteristics of the learning materials that may influence how bias is perceived. In this study, academic disciplines and material types were sourced directly from the original metadata of the OERs from which the sentences were extracted. These attributes provide additional context for interpreting sentences within their original environments, as context can affect what is perceived as biased (or non-biased). More details are provided in Table 4.

Table 4. Context features converted from categorical to numerical variables.

Category	Feature	Description
context	Discipline: Applied Sciences	$disc\_applied$ = *1 or 0*
context	Discipline: Arts and Humanities	$disc\_arts$ = *1 or 0*
context	Discipline: Mathematics	$disc\_math$ = *1 or 0*
context	Discipline: Natural Sciences	$disc\_natural$ = *1 or 0*
context	Discipline: Social Sciences	$disc\_social$ = *1 or 0*
context	Type: Activity	$type\_activ$ = *1 or 0*
context	Type: Course	$type\_course$ = *1 or 0*
context	Type: Lesson	$type\_lesson$ = *1 or 0*
context	Type: Other	$type\_other$ = *1 or 0*

3.3.1 Feature Expansion

To enhance the dataset and improve bias classification, we expanded the feature set by extracting linguistic, psycholinguistic, and contextual attributes from both sentences and their surrounding excerpts. This process involved three key steps: part-of-speech (PoS) tagging, computing psycholinguistic ratios, and determining sentiment valence scores.

1. **PoS Tagging:** PoS tagging was performed using the Stanford PoS Tagger via the NLTK library, selected for its accuracy in parsing text structure. This step identified word categories (e.g., nouns, verbs, adjectives), which were then used to compute linguistic abstraction features such as the adjective-to-verb ratio and sentence length.

2. **Psycholinguistic Feature Computation:** We extracted various psycholinguistic markers, including self-reference, socialization, and emotional stability, based on established linguistic models. These features may capture underlying biases in text by quantifying differences in writing styles and group references.

3. **Sentiment Analysis:** Sentiment valence scores were computed using VADER (Valence Aware Dictionary and sEntiment Reasoner) to determine the emotional tone of sentences. Scores ranged from -1 (negative) to $+1$ (positive), providing additional context on how ethnic groups are described.

The final feature set included both sentence-level and excerpt-level attributes, ensuring a broader contextual representation. This expansion aimed to improve the model's ability to detect subtle biases beyond isolated word occurrences, making bias classification more context aware.

3.3.2 Feature Selection

The feature selection process was based on a correlation-based filter approach, commonly used in exploratory ML studies (Guyon & Elisseeff, 2003; Chandrashekar & Sahin, 2014). Accordingly, Pearson's product-moment correlation test was applied to examine the linear relationship between each feature and the bias score, which captured participants' perceptions of ethnic bias at the sentence level (as detailed in Section 3.1). Features showing a statistically significant correlation ($p < 0.05$) with the bias score were retained for model development. To adjust for multiple comparisons, we applied the Benjamini–Hochberg False Discovery Rate correction using the *fdr_bh* option in the `multipletests` function from the `statsmodels` Python package (version 0.14.4). This implementation adjusts p-values and is consistent with the original procedure proposed by Benjamini and Hochberg (1995). While various feature selection methods exist, this correlation-based approach was chosen to maximize interpretability and transparency, supporting our goal of exploring which linguistic features might relate to student perceptions of bias.

Recognizing that Pearson correlation assumes linearity and normality, we also conducted a robustness check using Spearman's rank correlation, a non-parametric alternative that does not rely on these assumptions. The results were largely consistent, with only a few low-correlation features (four out of 17) not overlapping between the two methods. Based on this agreement and the interpretability of Pearson's method, the full Pearson-selected feature set was retained for model development.

In addition, to handle categorical features like discipline or type of educational material, we transformed the data using one-hot encoding with the `scikit-learn` Python library. This technique converted these categories into binary values, such as features *disc_applied* and *disc_arts* (see Table 4), creating a clear numerical representation of the data. This process ensured that the features were well structured and ready for correlation analysis, allowing us to assess the relationship between context and perceived bias.

3.4 Models

Identifying the best ML algorithm for detecting ethnic bias in educational texts is challenging, as most existing methods are designed for non-educational settings and often ignore contextual factors. To address this, we explored and compared several binary classification algorithms, including

- LG,
- decision tree (DT),
- NB,
- SVMs,
- K-nearest neighbours (KNN),
- RF, and
- extreme gradient boosting (XGB).

These algorithms were chosen for their effectiveness in binary classification tasks and their documented success in previous studies. For instance, RF and XGB, as ensemble models, were highlighted for their potential to improve performance by combining predictions from multiple trees (T. Chen & Guestrin, 2016; Sagi & Rokach, 2018). We also adopted a stacking (STK) strategy, which involved combining groups of two to seven models from the selected algorithms, resulting in 120 unique model configurations. Each combination was evaluated using the F1-score to identify the best-performing ensemble. The top model was then validated with an independent test dataset, ensuring its robustness and generalizability.

3.5 Evaluation Strategy

Below, we describe the steps taken to evaluate each model, including data transformations, evaluation metrics, and fine-tuning strategies.

3.5.1 Data Transformations

Before implementing each model, we prepared the data by scaling all features to a range of 0 to 1. This step aimed at standardizing all inputs across different algorithms and improving compatibility during model training.

Afterwards, we divided the dataset into two smaller sets: one contained 80% of samples to be used during the model development, and the other (20% of samples) was reserved for model evaluation. This typical 80/20 split is known for reducing the evaluation bias of the model's ability to generalize to unseen data. In addition, the data also had an uneven distribution of classes in which the "biased" label had 75 samples and the "not biased" label had 270 samples. This imbalance can skew the predictions toward the "not biased" class, given that certain ML algorithms struggle with discrepant class distributions (Chawla et al., 2002; Singhal et al., 2018). To prevent this, the minority class was oversampled using the synthetic minority over-sampling technique (SMOTE), a common approach for imbalanced datasets previously adopted in multiple domains (Chawla et al., 2002; Fernández et al., 2018).

Furthermore, to ensure fair representation of both classes across the training and validation sets, we also used the stratified K-fold cross-validation (SKCV) technique for training and validation. This technique divides the data into k folds and keeps the original class distribution in each fold. Thus, during each fold, the model was trained on $k-1$ folds and validated on the remaining fold. A notable mention is that the SMOTE strategy was only applied to the training data within each fold to prevent data leakage into the validation set. This was done to reduce the risk of overfitting while providing a reliable estimate of model performance in real-world scenarios (Berrar, 2019).

3.5.2 Metrics

Regarding evaluation metrics, we selected the F1-score as the primary performance metric. This combines precision and recall into a single value, making it particularly effective for tasks with imbalanced datasets. This choice was informed by F1's ability to balance the trade-off between precision (correctly predicting bias) and recall (capturing all actual instances of bias), which can provide a more reliable assessment of model performance. In addition, we also reported precision, recall, and accuracy so that we could have more insights about each model. Finally, to clarify class definitions, we defined the positive class as "biased" content and the negative class as "not biased" content.

3.5.3 Model Comparison

To assess the statistical robustness of differences in model performance, we applied a bootstrap resampling procedure with 10,000 iterations to the full dataset. In each iteration, training data were resampled with replacement, ensuring class proportions were maintained, followed by SMOTE oversampling to address class imbalance. Accordingly, models were fitted, and performance metrics (F1-score, accuracy) were evaluated on the out-of-bag (OOB) samples. For each model pair, we computed the distribution of metric differences across resamples, estimated the 95% confidence interval using the 2.5th and 97.5th percentiles, and calculated the mean difference. Statistical significance of the pairwise differences between model performance distributions was additionally assessed using a two-sided Wilcoxon signed-rank test. This approach aligns with recent methodological standards in LA that emphasize quantifying uncertainty in model evaluation (Choi et al., 2025; Borchers & Baker, 2025).

3.6 Fine-Tuning

To enhance performance, models were initially trained using default hyperparameters and then fine-tuned with the F1-score as the optimization target. Additionally, the fine-tuning process was conducted using a random search strategy, which selects random combinations of hyperparameters from a predefined grid. This approach was chosen for being more time efficient than exhaustive grid search while still exploring a wide range of configurations (Bergstra & Bengio, 2012).

The ranges of hyperparameters tested for each classifier are illustrated in Table 5, which shows the range of configurations explored during the fine-tuning process.

4. Results

This section presents the results of our analysis. We begin by examining the relevance of selected features in detecting ethnic bias in learning materials (Section 4.1), followed by an evaluation of the classification models (Section 4.2).

4.1 Feature Analysis

Below we present the answer to *RQ1: Which features might support the identification of ethnic bias in text-based online learning materials?*, which focused on potential features for identifying ethnic bias in learning texts. Accordingly, the correlation analysis between ethnic bias and expanded features showed a range of statistically significant correlations between certain features and the bias score. Such correlations ranged from negative to positive and are detailed in Table 6.

The columns in Table 6 are defined as follows: *category* refers to the feature category (e.g., contextual attributes, linguistic abstraction), *feature* is the specific feature identifier, t is the t-statistic from the correlation test, *p-value* is the respective p-value

Table 5. Range of hyperparameters tested for each classifier.

Hyperparameter	Range tested	LG	NB	SVM	KNN	DT	RF	XGB
solver	['newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga']	x						
C	np.logspace(2, -2, num = 30)	x		x				
gamma	np.logspace(0, -4, num = 10)			x				x
kernel	['linear', 'poli', 'rbf', 'sigmoid']			x				
var_smoothing	np.logspace(0, -10, num = 100)		x					
leaf_size	[1..10]				x			
n_neighbors	[1..10]				x			
p	[1, 2]				x			
algorithm	['ball_tree', 'kd_tree', 'brute']				x			
criterion	['gini', 'entropy', 'log_loss']					x		
max_depth	[2..50]					x	x	x
min_samples_split	[5..15]						x	
n_estimators	[10, 20, 30, ..., 700]						x	x
bootstrap	[True, False]						x	
max_features	[1..16]					x	x	
min_samples_leaf	[1..10]						x	
learning_rate	np.logspace(-1, -4, num = 10)							x
min_child_weight	[1..5]							x

*For other hyperparameters, their default values were kept.

from the correlation test, r indicates the Pearson correlation coefficient, and *95% CI* shows the confidence interval of r, with *lower* and *upper* as the lower and upper bounds, respectively.

Notable features include *target_ethnic* ($r = -0.173$, $p = 0.001$) and *disc_arts* ($r = -0.159$, $p = 0.003$), which showed negative correlations with the bias score. Other features, such as *vrb_ratio* and *aggressiv*, showed positive correlations ($r = 0.193$, < 0.001). Context-based features extracted from the surrounding text (e.g., *self_ref_exc*, $r = -0.158$, $p = 0.003$) were also statistically significantly correlated with the bias score.

Overall, features with statistically significant correlations ($p < 0.05$) were selected for the classification task. These include attributes from multiple categories, such as linguistic abstraction (*vrb_ratio*), psycholinguistic markers (*aggressiv*), and contextual attributes (*disc_arts*). These results suggest potential features for training classification models for identifying ethnic bias in text-based learning materials. The next section details the performance of classification models trained using these identified features.

4.2 Model Performance

This section presents the results of model training, fine-tuning, and testing.

4.2.1 Training and Fine-Tuning

Post fine-tuning, the optimal hyperparameters for each model were identified and are listed in Table 7. Only the optimized hyperparameters are included (i.e., hyperparameters with default values are omitted for brevity).

In the STK experiment, which evaluated 120 possible model combinations, the best-performing stack achieved an F1-score of 0.88. This stack included LG, SVM, NB, KNN, and XGB models and outperformed individual models such as SVM, which achieved the second-highest F1-score of 0.86. A summary of the fine-tuned model performances is presented in Figure 3.

4.2.2 Testing

After fine-tuning, the models were evaluated on a separate test dataset to assess their performance in identifying ethnic bias in unseen data (*RQ2: Which classification approaches might be suitable for identifying ethnic bias in text-based online learning materials?*). Table 8 summarizes the results for each model on this specific test set, including precision, recall, F1-score, and accuracy. Both individual models and the best-performing stack were tested, with results calculated for classifying texts as "biased" or "not biased."

Table 8 shows that on this test set, the stacked model (STK) and the SVM model both achieved the highest accuracy (0.75), with similar performance in terms of F1-score (0.71 for SVM vs. 0.69 for STK). RF and XGB models also performed comparably, with F1-scores of 0.70 and 0.69, respectively. In contrast, the LG model exhibited lower precision and recall for the "biased" class, resulting in an overall accuracy of 0.58.

Table 6. Correlations between potential features and ethnic bias. Statistically significant p-values are highlighted in boldface. Rows are sorted by the Pearson correlation coefficient (r).

						95% CI	
category	**feature***	**t**	**p-value**	**Adj. p-value (BH)**	**r**	**lower**	**upper**
mentions	target_ethnic	-3.257	**0.001**	**0.010**	-0.173	-0.274	-0.069
context	disc_arts	-2.976	**0.003**	**0.015**	-0.159	-0.260	-0.054
psychological	self_ref_exc	-2.957	**0.003**	**0.015**	-0.158	-0.259	-0.053
context	type_lesson	-2.628	**0.009**	**0.033**	-0.140	-0.242	-0.035
mentions	ethnic_maj	-2.516	**0.012**	**0.037**	-0.135	-0.237	-0.029
abstraction	adj_ratio_exc	-2.135	**0.033**	0.094	-0.115	-0.217	-0.009
psychological	self_ref	-1.931	0.054	0.127	-0.104	-0.207	0.002
abstraction	adj_ratio	-1.698	0.090	0.198	-0.091	-0.195	0.014
abstraction	sent_len	-1.555	0.121	0.242	-0.084	-0.188	0.022
context	type_other	-1.265	0.207	0.345	-0.068	-0.172	0.038
valence	valence_exc	-0.882	0.379	0.561	-0.048	-0.152	0.058
abstraction	pnoun_ratio_exc	-0.660	0.509	0.702	-0.036	-0.141	0.070
valence	valence	-0.594	0.553	0.722	-0.032	-0.137	0.074
mentions	ethnic_maj_exc	-0.507	0.612	0.722	-0.027	-0.133	0.078
abstraction	sent_len_exc	-0.505	0.614	0.722	-0.027	-0.132	0.079
abstraction	noun_ratio	-0.285	0.776	0.862	-0.015	-0.121	0.090
context	disc_math	-0.132	0.895	0.933	-0.007	-0.113	0.099
context	disc_natural	-0.132	0.895	0.933	-0.007	-0.113	0.099
context	type_course	-0.113	0.910	0.933	-0.006	-0.112	0.100
abstraction	noun_ratio_exc	-0.031	0.976	0.976	-0.002	-0.107	0.104
context	disc_applied	0.386	0.699	0.799	0.021	-0.085	0.126
mentions	ethnic_min_exc	0.515	0.607	0.722	0.028	-0.078	0.133
abstraction	unique_pos_exc	0.568	0.571	0.722	0.031	-0.075	0.136
abstraction	adv_ratio_exc	0.758	0.449	0.641	0.041	-0.065	0.146
abstraction	pnoun_ratio	0.949	0.344	0.529	0.051	-0.055	0.156
abstraction	adv_ratio	0.952	0.342	0.529	0.051	-0.055	0.156
psychological	socialisation_exc	1.327	0.186	0.323	0.071	-0.034	0.176
psychological	emo_stability	1.382	0.168	0.305	0.074	-0.031	0.179
abstraction	unique_word_exc	1.386	0.167	0.305	0.075	-0.031	0.179
mentions	ethnic_min	1.677	0.094	0.198	0.090	-0.016	0.194
abstraction	vrb_ratio_exc	2.063	**0.040**	0.100	0.111	0.005	0.214
psychological	aggressiv_exc	2.063	**0.040**	0.100	0.111	0.005	0.214
psychological	socialisation	2.521	**0.012**	**0.037**	0.135	0.030	0.237
psychological	emo_stability_exc	2.780	**0.006**	**0.024**	0.148	0.044	0.250
abstraction	unique_word	2.807	**0.005**	**0.022**	0.150	0.045	0.251
abstraction	unique_pos	2.993	**0.003**	**0.015**	0.160	0.055	0.261
context	disc_social	3.069	**0.002**	**0.015**	0.163	0.059	0.264
abstraction	vrb_ratio	3.647	**<0.001**	**0.010**	0.193	0.089	0.293
psychological	aggressiv	3.647	**<0.001**	**0.010**	0.193	0.089	0.293
context	type_activ	3.854	**<0.001**	**0.010**	0.204	0.100	0.303

* Features ending in "_exc" are the ones extracted from the excerpt (i.e., the text surrounding each sentence).

Table 7. Hyperparameters that led to the best F1-score after fine-tuning the models using the random search strategy.

Model	**Hyperparameters**
LG	C = 0.452, solver = 'newton-cg'
SVM	C = 4.641, gamma = 0.360, probability = True
NB	var_smoothing = 0.792
KNN	algorithm = 'ball_tree', leaf_size = 8, n_neighbors = 1, p = 1
DT	criterion = 'entropy', max_depth = 14, max_features = 7
RF	max_depth = 48, max_features = 2, min_samples_split = 5, n_estimators = 250
XGB	objective = 'binary:logistic', booster = 'gbtree', learning_rate = 0.1, n_estimators = 290, max_depth = 14, min_child_weight = 3, gamma = 0.002

4.2.3 Pairwise Model Comparisons

Table 9 presents the top differences in accuracy and F1-score, sorted by absolute mean difference and filtered to include only statistically significant results ($p < 0.05$). This selection highlights the most substantial observed differences while maintaining statistical rigour.

Although many comparisons yielded very small p-values (due to the high number of bootstrap iterations), not all were accompanied by confidence intervals excluding zero. Accordingly, we interpret these findings as indicative of performance tendencies rather than definitive evidence of superiority.

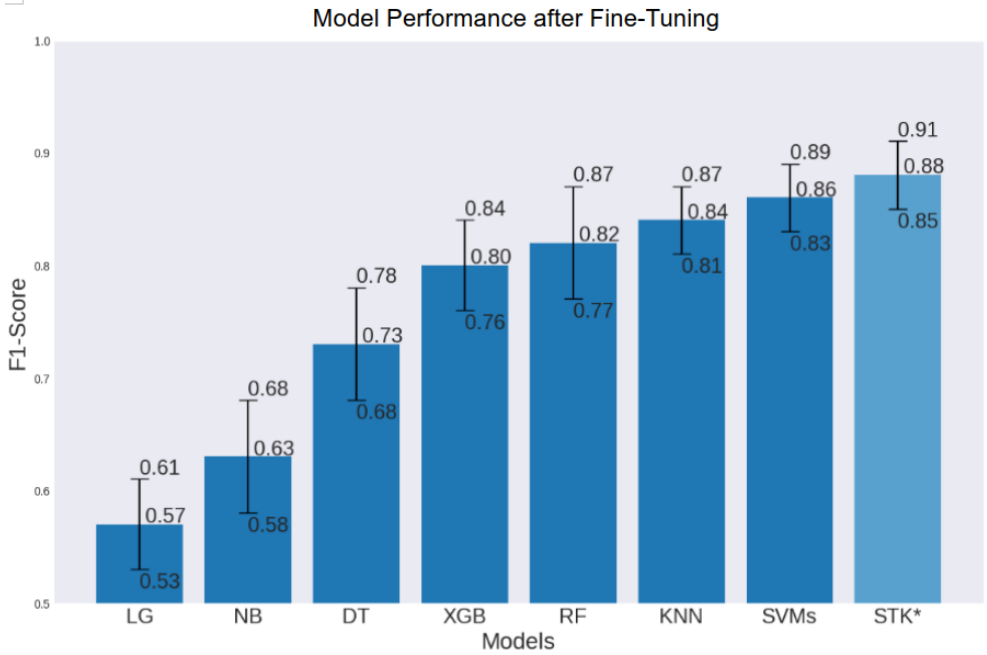

Figure 3. Model performance across different classifiers after fine-tuning. Bars represent mean F1-scores averaged over five-fold stratified cross-validation; error bars represent standard deviation across folds.

*STK = The stack leading to the best performance (models stacked: LG, SVM, NB, KNN, and XGB).

Table 8. Models' final performance after being tested against the test data. "Average" refers to the weighted average based on the size of each class.

		Precision	Recall	F1-score	n	Accuracy
LG	Not Biased	0.82	0.59	0.69	54	0.58
	Biased	0.27	0.53	0.36	15	
	Average	0.70	0.58	0.62	69	
SVM	Not Biased	0.79	0.93	0.85	54	0.75
	Biased	0.33	0.13	0.19	15	
	Average	0.69	0.75	0.71	69	
NB	Not Biased	0.88	0.41	0.56	54	0.49
	Biased	0.27	0.80	0.41	15	
	Average	0.75	0.49	0.52	69	
KNN	Not Biased	0.79	0.76	0.77	54	0.65
	Biased	0.24	0.27	0.25	15	
	Average	0.67	0.65	0.66	69	
DT	Not Biased	0.80	0.59	0.68	54	0.57
	Biased	0.24	0.47	0.32	15	
	Average	0.68	0.57	0.60	69	
RF	Not Biased	0.80	0.83	0.82	54	0.71
	Biased	0.31	0.27	0.29	15	
	Average	0.70	0.71	0.70	69	
XGB	Not Biased	0.80	0.81	0.81	54	0.70
	Biased	0.29	0.27	0.28	15	
	Average	0.69	0.70	0.69	69	
STK*	Not Biased	0.78	0.94	0.86	54	0.75
	Biased	0.25	0.07	0.11	15	
	Average	0.67	0.75	0.69	69	

*STK = The stack including the models LG, SVM, NB, KNN, and XGB.

Below, we discuss these results and their implications for future research in LA and artificial intelligence in education.

5. Discussion

In this study, we explored traditional ML models to identify ethnic bias in online learning texts. In particular, we addressed two research questions: *RQ1: Which features might support the identification of ethnic bias in text-based online learning materials?* and *RQ2: Which classification approaches might be suitable for identifying ethnic bias in text-based online learning materials?*.

Table 9. Top 20 pairwise differences in accuracy and F1-score from bootstrap analysis.

Metric	Model A	Model B	Mean Diff (A-B)	95% CI	p-value
Accuracy	LG	RF	-0.134	[-0.256, -0.017]	<0.01
Accuracy	LG	STK	-0.119	[-0.246, 0.000]	<0.01
Accuracy	LG	XGB	-0.116	[-0.240, 0.000]	<0.01
Accuracy	KNN	RF	-0.108	[-0.195, -0.024]	<0.01
Accuracy	DT	RF	-0.103	[-0.202, -0.008]	<0.01
Accuracy	KNN	STK	-0.093	[-0.165, -0.025]	<0.01
Accuracy	KNN	XGB	-0.089	[-0.180, 0.000]	<0.01
Accuracy	NB	RF	-0.089	[-0.292, 0.041]	<0.01
Accuracy	DT	STK	-0.088	[-0.190, 0.016]	<0.01
Accuracy	DT	XGB	-0.085	[-0.189, 0.016]	<0.01
Accuracy	NB	STK	-0.074	[-0.282, 0.065]	<0.01
Accuracy	NB	XGB	-0.071	[-0.279, 0.068]	<0.01
F1-Score	NB	STK	0.069	[-0.116, 0.259]	<0.01
Accuracy	SVC	RF	-0.068	[-0.169, 0.023]	<0.01
Accuracy	LG	SVC	-0.066	[-0.189, 0.049]	<0.01
F1-Score	LG	STK	0.062	[-0.109, 0.242]	<0.01
F1-Score	NB	KNN	0.060	[-0.107, 0.226]	<0.01
F1-Score	NB	RF	0.055	[-0.120, 0.240]	<0.01
F1-Score	LG	KNN	0.054	[-0.099, 0.208]	<0.01
F1-Score	NB	DT	0.054	[-0.120, 0.226]	<0.01

(Showing top 20 rows, sorted by absolute mean difference.)

Using a dataset of 345 sentences labelled by diverse ethnic groups, we identified features potentially linked to ethnic bias in text-based materials. Then, several ML models—including ensemble strategies—were applied to classify bias. The bootstrap analysis indicated that RF and STK consistently showed higher performance in distinguishing biased from non-biased content, while the NB model demonstrated the highest precision. (RF achieved an F1-score of 0.70 on the test set, and NB achieved 0.75 precision on the same test set.) These findings highlight the potential for automated tools to detect bias in educational materials and lay the groundwork for developing fairer and more inclusive learning technologies. Below, we discuss key findings and implications, followed by limitations and directions for future research and practice.

5.1 Features

One key finding is the statistically significant correlations between certain features and ethnic bias in online learning texts. These features covered categories such as linguistic abstraction (*vrb_ratio*, *adj_ratio*), psycholinguistic markers (*aggressiv*), and contextual attributes (*disc_arts*, *type_lesson*). This suggests that subtle elements of language, such as the use of abstract terms or references to specific disciplines, can signal potential ethnic bias in educational materials. Previous research on linguistic intergroup bias (Maass, 1999) has demonstrated that abstract language can reflect perceptions of in-group and out-group members. Similarly, psycholinguistic markers have been linked to emotional states that influence social group dynamics (Sboev et al., 2015; Tajfel, 1970). However, these studies have primarily focused on generic settings rather than ethnic bias in learning contexts. By applying these concepts to online learning materials, this study provides new insights into how linguistic and contextual features can reveal ethnic bias, extending existing research.

Another key finding is features related to social sciences, such as *disc_arts* and *type_lesson*, which showed a strong connection to perceived bias. One possible explanation is that social sciences often engage directly with topics like ethnicity and intergroup dynamics, where bias may be more explicitly discussed or inferred. In contrast, disciplines like mathematics and physics rarely address such themes, potentially making bias less visible in their content. Accordingly, these findings highlight the need to address discipline-specific biases in learning content, which aligns with previous studies regarding the presence of bias in particular curricula (Skopec et al., 2021). However, it is important to note that this does not imply that social sciences inherently promote bias but rather reflects the nature of their subject matter. Educators and institutions should focus on carefully reviewing content in each field to ensure fair and inclusive ethnic representations.

We also found that contextual features, such as *self_ref_exc*, were potential indicators of bias. This highlights the importance of analyzing not just individual sentences but also the surrounding text, as context can provide critical clues about the subjective

nature of bias. Previous studies have shown that perceptions of bias are often shaped by cultural and moral judgments (Brewer & Yuki, 2007; Chiu et al., 1997). For instance, the same phrase might be interpreted differently depending on the cultural context or the broader narrative in which it appears. These findings suggest that automated tools for bias detection must account for contextual information to provide a more accurate and nuanced understanding of bias.

Overall, these findings show that specific linguistic and contextual features, such as *vrb_ratio*, *disc_arts*, and *self_ref_exc*, can help identify ethnic bias in learning materials. For example, *disc_arts* points to biases that may arise in social sciences, where topics like ethnicity and intergroup dynamics are often discussed. Contextual features like *self_ref_exc* highlight the importance of analyzing the surrounding text, as biases may not always be clear in single sentences. This study moves beyond earlier research by focusing on text-based learning materials rather than general text or word-level biases. In summary, these results can support educators and learning designers in spotting specific features that indicate bias, helping to create fairer and more inclusive learning environments.

5.2 ML Models

The findings from our bootstrap analysis indicate that RF and STK were among the more effective models for detecting ethnic bias in online learning materials, consistently achieving competitive F1-scores and accuracy (Table 9). Specifically, while the point estimates for accuracy showed SVM and STK with the highest performance on the test set (Table 8), the bootstrap comparisons revealed more nuanced relationships. For instance, RF exhibited a robustly higher accuracy than LG, KNN, and DT, with 95% confidence intervals for the differences entirely excluding zero. The STK approach—which combined LG, SVM, NB, KNN, and XGB—often showed performance similar to or slightly better than individual base models in the bootstrap distribution, reinforcing the potential of ensemble strategies for consistent performance.

In contrast, NB consistently struggled with recall (0.49 on the test set) and accuracy (0.49), limiting its ability to detect all biased instances in this study. However, it achieved notably higher precision on the test set (0.75), making it valuable for applications where minimizing false positives is the priority. More broadly, these findings, especially those robustly supported by bootstrap confidence intervals, align with previous research on ML in bias detection, which highlights the need for context-specific model selection based on performance priorities (Q. Nguyen et al., 2020; Baker & Hawn, 2022; Pryzant et al., 2020).

5.2.1 Practical Implications

The results provide initial indications of how ML models might enhance bias detection workflows for educators, curriculum designers, and learning material reviewers. Specifically, these models could be used to generate outputs such as bias scores, flagged sentences, or ranked lists of content items based on their likelihood of containing ethnic bias. These analytics may help educators prioritize which materials to review manually, focus attention on higher-risk areas, and identify broader patterns (e.g., which disciplines or resource types are more frequently flagged).

Based on the observed performance and the bootstrap analysis, RF models appear particularly suitable for tasks where balanced performance (F1-score) and higher accuracy are important, such as pre-screening educational materials for potential biases, given its robust significant differences over several simpler models. The STK approach, by combining multiple models, offers relatively balanced performance across metrics and could assist in contexts where both recall and precision are valued. In contrast, NB may be useful for applications that prioritize minimizing false positives, such as reviewing flagged content from new datasets. Together, these models may support the content evaluation process, potentially reducing the manual workload and allowing educators to focus their efforts on improving content quality.

Figure 4 presents an illustrative decision pathway showing how different models might be selected based on specific educational priorities.

Beyond supporting educators, these findings may also have a positive impact on students by helping to foster more equitable learning environments. Automated pre-screening using models like RF could assist in identifying learning materials that are more inclusive and less likely to perpetuate biases. The balanced approach of stacking models may help reduce the likelihood of biased content slipping through reviews, supporting a more consistent learning experience. Additionally, bias detection using NB could support efforts to regularly update new materials to maintain fairness. Overall, these applications have the potential to contribute to a more inclusive and supportive educational experience.

Overall, choosing the appropriate model may also allow educators and course designers to apply AI-based tools more effectively, making bias detection efforts more focused and scalable (Skopec et al., 2021; Maass, 1999). For students, these automated approaches could promote more inclusive learning materials, contributing to a fairer educational environment. However, while these findings provide valuable insights, they also highlight constraints that must be addressed in future work. The next section discusses these limitations and potential areas for improvement.

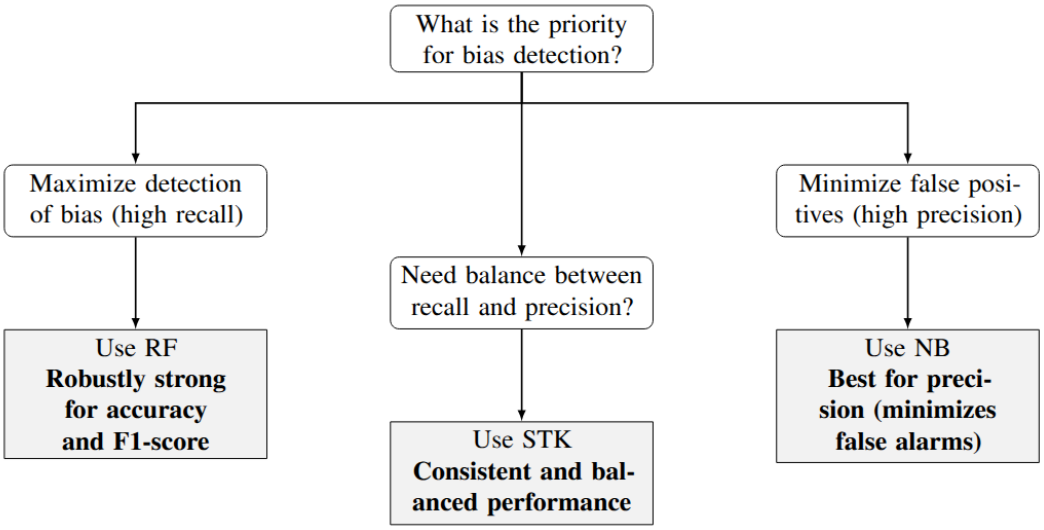

Figure 4. *Illustrative decision tree for selecting ML models for bias detection.*

5.3 Limitations and Future Directions

As with any other research, this study has several limitations which future research might consider addressing. First, the dataset was primarily composed of text excerpts evaluated by higher education students in the US and the UK. As a result, the findings may not fully generalize to other educational contexts, including different cultural, linguistic, or institutional settings. Future research should expand the dataset to include a broader range of perspectives and diverse learning materials to improve applicability.

Second, the study was based on a binary classification system to label sentences as either "biased" or "not biased." While this simplifies bias detection, it may not account for subtler variations or degrees of bias that may exist within learning materials. A more nuanced classification approach—such as ranking bias intensity or identifying different types of bias—could improve accuracy and offer deeper insights for content review.

Third, while traditional ML models such as RF and STK demonstrated robust performance in our bootstrap analysis, they still depend on pre-defined features for classification. More advanced techniques, such as deep learning and LLMs, could capture contextual and implicit biases more effectively. However, such methods would require larger datasets and careful evaluation to avoid perpetuating biases present in data training. In addition, although traditional models offer greater interpretability than deep learning architectures, ensemble methods like RF may still benefit from post hoc techniques (e.g., feature importance, SHAP) to enhance explainability, which future work could explore.

Finally, this study focused on text-based bias detection, but bias in educational materials extends beyond textual content to images, videos, and multimedia resources. Future research should explore multimodal approaches that integrate text analysis with other media types to provide a more comprehensive assessment of bias in digital learning environments.

Addressing these limitations may enhance the accuracy, fairness, and generalizability of AI-driven bias detection tools, ultimately supporting the development of more inclusive and equitable learning materials.

6. Conclusion

This study investigated the use of LA and ML to detect ethnic bias in online learning materials. By combining feature analysis with multiple classification models, it identified linguistic and contextual markers associated with bias and demonstrated the viability of automated approaches for bias detection in educational texts.

The results highlight the potential of AI-driven tools to support bias identification in learning materials. Specifically, while SVM showed strong point-estimate performance on the test set, the bootstrap analysis indicated that RF and STK consistently exhibited robust performance in distinguishing biased from non-biased content. The findings also underscore the importance of discipline-specific considerations, as social sciences content appeared more frequently associated with perceived bias in this sample.

While limitations exist, such as dataset scope and the use of a binary classification system, this study provides an initial base for further work in this area. Future research can expand dataset diversity, explore more nuanced bias detection methods, and incorporate advanced AI techniques, such as deep learning and multimodal analysis, to enhance accuracy and applicability. By

advancing research on automated bias detection, this study contributes to efforts toward developing more inclusive educational environments. As online learning continues to grow, ensuring fairness and representation in learning content remains a critical challenge—one that LA and AI-based methods are increasingly positioned to help address.

Declaration of Conflict of Interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The publication of this article received financial support from The Open University and the UK's Economic and Social Research Council (ESRC) (ES/Z504439/1).

References

Albuquerque, J. (2023). *Towards an automatic approach for uncovering ethnic bias in online learning texts* [Doctoral dissertation, The Open University]. https://doi.org/10.21954/ou.ro.000170d9

Albuquerque, J., Bittencourt, I. I., Coelho, J. A. P. M., & Silva, A. P. (2017). Does gender stereotype threat in gamified educational environments cause anxiety? An experimental study. *Computers & Education*, *115*, 161–170. https://doi.org/10.1016/j.compedu.2017.08.005

Albuquerque, J., Rienties, B., Holmes, W., & Hlosta, M. (2025). From hype to evidence: Exploring large language models for inter-group bias classification in higher education. *Interactive Learning Environments*, *33*(3), 2332–2354. https://doi.org/10.1080/10494820.2024.2408554

Allen, I. E., & Seaman, J. (2007). *Online nation: Five years of growth in online learning.* ERIC. https://files.eric.ed.gov/fulltext/ED529699.pdf

Al-Zawqari, A., & Vandersteen, G. (2023). Fairness in predictive learning analytics: A case study in online STEM education. In *Proceedings of the 2023 IEEE Frontiers in Education Conference* (FIE 2023), 18–21 October 2023, College Station, Texas, USA (pp. 1–5). IEEE. https://doi.org/10.1109/FIE58773.2023.10343059

Ashong, C. Y., & Commander, N. E. (2012). Ethnicity, gender, and perceptions of online learning in higher education. *MERLOT Journal of Online Learning and Teaching*, *8*(2), 98–110. https://jolt.merlot.org/vol8no2/ashong_0612.pdf

Baker, R. S., & Hawn, A. (2022). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, *32*(4), 1052–1092. https://doi.org/10.1007/s40593-021-00285-9

Balica, R. (2018). Big data learning analytics and algorithmic decision-making in digital education governance. *Analysis and Metaphysics*, *17*, 128–133. https://doi.org/10.22381/AM1720187

Bansak, C., & Starr, M. (2021). Covid-19 shocks to education supply: How 200,000 US households dealt with the sudden shift to distance learning. *Review of Economics of the Household*, *19*(1), 63–90. https://doi.org/10.1007/s11150-020-09540-9

Baumeister, R., & Vohs, K. (2007). Ingroup–outgroup bias. In *Encyclopedia of social psychology* (pp. 484–485). SAGE Publications, Inc. https://doi.org/10.4135/9781412956253.n286

Beaunoyer, E., Dupéré, S., & Guitton, M. J. (2020). COVID-19 and digital inequalities: Reciprocal impacts and mitigation strategies. *Computers in Human Behavior*, *111*, 106424. https://doi.org/10.1016/j.chb.2020.106424

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, *13*(2), 281–305. https://www.jmlr.org/papers/volume13/bergstra12a/bergstra12a.pdf

Berrar, D. (2019). Cross-validation. In S. Ranganathan, M. Gribskov, K. Nakai, & C. Schönbach (Eds.), *Encyclopedia of bioinformatics and computational biology* (pp. 542–545, Vol. 1). Academic Press. https://doi.org/10.1016/B978-0-12-809633-8.20349-X

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., … Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*. https://doi.org/10.48550/arXiv.2108.07258

Borchers, C., & Baker, R. S. (2025). ABROCA distributions for algorithmic bias assessment: Considerations around interpretation. In *Proceedings of the 15th International Conference on Learning Analytics and Knowledge* (LAK 2025), 3–7 March 2025, Dublin, Ireland (pp. 837–843). ACM. https://doi.org/10.1145/3706468.3706498

Brewer, M., & Yuki, M. (2007). Culture and social identity [PsycINFO ID: 2007-12976-012]. In S. Kitayama & D. Cohen (Eds.), *Handbook of cultural psychology* (1st ed., pp. 307–322). The Guilford Press.

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, *40*(1), 16–28. https://doi.org/10.1016/j.compeleceng.2013.11.024

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Chen, F., & Cui, Y. (2020). Utilizing student time series behaviour in learning management systems for early prediction of course performance. *Journal of Learning Analytics*, *7*(2), 1–17. https://doi.org/10.18608/jla.2020.72.1

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD 2016), 13–17 August 2016, San Francisco, California, USA (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785

Chiu, C.-y., Hong, Y.-y., & Dweck, C. S. (1997). Lay dispositionism and implicit theories of personality. *Journal of Personality and Social Psychology*, *73*(1), 19–30. https://doi.org/10.1037//0022-3514.73.1.19

Choi, J., Karumbaiah, S., & Matayoshi, J. (2025). Bias or insufficient sample size? Improving reliable estimation of algorithmic bias for minority groups. In *Proceedings of the 15th International Conference on Learning Analytics and Knowledge* (LAK 2025), 3–7 March 2025, Dublin, Ireland (pp. 547–557). ACM. https://doi.org/10.1145/3706468.3706540

Copur-Gencturk, Y., Thacker, I., & Cimpian, J. R. (2022). Teacher bias in the virtual classroom. *Computers & Education*, *191*, 104627. https://doi.org/10.1016/j.compedu.2022.104627

Dai, S., Xu, C., Xu, S., Pang, L., Dong, Z., & Xu, J. (2024). Bias and unfairness in information retrieval systems: New challenges in the LLM era. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (KDD 2024), 25–29 August 2024, Barcelona, Spain (pp. 6437–6447). ACM. https://doi.org/10.1145/3637528.3671458

Daumeyer, N. M., Onyeador, I. N., Brown, X., & Richeson, J. A. (2019). Consequences of attributing discrimination to implicit vs. explicit bias. *Journal of Experimental Social Psychology*, *84*, 103812. https://doi.org/10.1016/j.jesp.2019.04.010

Dawkins, H., Hedgeland, H., & Jordan, S. (2017). Impact of scaffolding and question structure on the gender gap. *Physical Review Physics Education Research*, *13*(2), 020117. https://doi.org/10.1103/PhysRevPhysEducRes.13.020117

Doubé, W., & Lang, C. (2012). Gender and stereotypes in motivation to study computer programming for careers in multimedia. *Computer Science Education*, *22*(1), 63–78. https://doi.org/10.1080/08993408.2012.666038

Dovidio, J. F., & Gaertner, S. L. (2010). Intergroup bias. In *Handbook of social psychology* (pp. 1084–1121). John Wiley & Sons, Inc. https://doi.org/10.1002/9780470561119.socpsy002029

Dutt, A., Ismail, M. A., & Herawan, T. (2017). A systematic review on educational data mining. *IEEE Access*, *5*, 15991–16005. https://doi.org/10.1109/ACCESS.2017.2654247

Ferguson, R. (2012). Learning analytics: Drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, *4*(5/6), 304–317. https://doi.org/10.1504/IJTEL.2012.051816

Fernández, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. *Journal of Artificial Intelligence Research*, *61*, 863–905. https://doi.org/10.1613/jair.1.11192

Greenwald, A. G., & Krieger, L. H. (2006). Implicit bias: Scientific foundations. *California Law Review*, *94*(4), 945–967. https://doi.org/10.2307/20439056

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*(Mar), 1157–1182. https://dl.acm.org/doi/10.5555/944919.944968

Hollister, B., Nair, P., Hill-Lindsay, S., & Chukoskie, L. (2022). Engagement in online learning: Student attitudes and behavior during COVID-19. *Frontiers in Education*, *7*, 851019. https://doi.org/10.3389/feduc.2022.851019

Holstein, K., & Doroudi, S. (2019). Fairness and equity in learning analytics systems (FairLAK). In *Workshop at the Ninth International Conference on Learning Analytics and Knowledge* (LAK 2019), 4–8 March 2019, Tempe, Arizona, USA. https://sites.google.com/view/fairlak

Holstein, K., & Doroudi, S. (2022). Equity and artificial intelligence in education. In W. Holmes & K. Porayska-Pomsta (Eds.), *The ethics of artificial intelligence in education* (pp. 151–173). Routledge. https://www.taylorfrancis.com/chapters/edit/10.4324/9780429329067-9/equity-artificial-intelligence-education-kenneth-holstein-shayan-doroudi

Hutt, S., Baker, R. S., Ashenafi, M. M., Andres-Bray, J. M., & Brooks, C. (2022). Controlled outputs, full data: A privacy-protecting infrastructure for mooc data. *British Journal of Educational Technology*, *53*(4), 756–775. https://doi.org/10.1111/bjet.13231

Jordano, M. L., & Touron, D. R. (2017). Stereotype threat as a trigger of mind-wandering in older adults. *Psychology and Aging*, *32*(3), 307. https://doi.org/10.1037/pag0000167

Karumbaiah, S., & Brooks, J. (2021). How colonial continuities underlie algorithmic injustices in education. In C. Gardner-McCune, S. Grady, Y. Jimenez, J. Ryoo, R. Santo, & J. Payton (Eds.), *Proceedings of the 2021 Conference on Research in Equitable and Sustained Participation in Engineering, Computing, and Technology* (RESPECT 2021), 23–27 May 2021, online (pp. 1–6). IEEE. https://doi.org/10.1109/RESPECT51740.2021

Kizilcec, R. F., & Lee, H. (2022). Algorithmic fairness in education. In W. Holmes & K. Porayska-Pomsta (Eds.), *The ethics of artificial intelligence in education* (pp. 174–202). Routledge. https://www.taylorfrancis.com/chapters/edit/10.4324/9780429329067-10/algorithmic-fairness-education-ren

Kumar, D., Jain, U., Agarwal, S., & Harshangi, P. (2024). Investigating implicit bias in large language models: A large-scale study of over 50 LLMs. *arXiv preprint arXiv:2410.12864*. https://doi.org/10.48550/arXiv.2410.12864

Liao, Q. V., & Varshney, K. R. (2021). Human-centered explainable AI (XAI): From algorithms to user experiences. *arXiv preprint arXiv:2110.10790*. https://doi.org/10.48550/arXiv.2110.10790

Long, P., Siemens, G., Gráinne, C., & Gašević, D. (2011). 1st International Conference on Learning Analytics and Knowledge. In *Proceedings of the First International Conference on Learning Analytics and Knowledge* (LAK 2011), 27 February–1 March 2011, Banff, Alberta, Canada (pp. 3–4). ACM. https://dl.acm.org/doi/proceedings/10.1145/2090116

Maass, A. (1999). Linguistic intergroup bias: Stereotype perpetuation through language. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (pp. 79–121, Vol. 31). Elsevier. https://doi.org/10.1016/S0065-2601(08)60272-5

Mayer, R. E. (2019). How multimedia can improve learning and instruction. In J. Dunlosky & K. A. Rawson (Eds.), *The Cambridge handbook of cognition and education* (pp. 460–479). Cambridge University Press. https://doi.org/10.1017/9781108235631.019

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, *54*(6), 1–35. https://doi.org/10.1145/3457607

Mendelsohn, J., Tsvetkov, Y., & Jurafsky, D. (2020). A framework for the computational linguistic analysis of dehumanization. *Frontiers in Artificial Intelligence*, *3*, 55. https://doi.org/10.3389/frai.2020.00055

Mohamed, S., Png, M.-T., & Isaac, W. (2020). Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology*, *33*, 659–684. https://doi.org/10.1007/s13347-020-00405-8

Namoun, A., & Alshanqiti, A. (2020). Predicting student performance using data mining and learning analytics techniques: A systematic literature review. *Applied Sciences*, *11*(1), 237. https://doi.org/10.3390/app11010237

Nguyen, A., Ngo, H. N., Hong, Y., Dang, B., & Nguyen, B.-P. T. (2023). Ethical principles for artificial intelligence in education. *Education and Information Technologies*, *28*(4), 4221–4241. https://doi.org/10.1007/s10639-022-11316-w

Nguyen, Q., Rienties, B., & Richardson, J. T. E. (2020). Learning analytics to uncover inequality in behavioural engagement and academic attainment in a distance learning setting. *Assessment & Evaluation in Higher Education*, *45*(4), 594–606. https://doi.org/10.1080/02602938.2019.1679088

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning*. University of Illinois Press.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. https://jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf

Pennington, C. R., Heim, D., Levy, A. R., & Larkin, D. T. (2016). Twenty years of stereotype threat research: A review of psychological mediators. *PLoS ONE*, *11*(1), 1–25. https://doi.org/10.1371/journal.pone.0146487

Pryzant, R., Martinez, R. D., Dass, N., Kurohashi, S., Jurafsky, D., & Yang, D. (2020). Automatically neutralizing subjective bias in text. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(01), 480–489. https://doi.org/10.1609/aaai.v34i01.5385

Richardson, J. T. E., Mittelmeier, J., & Rienties, B. (2020). The role of gender, social class and ethnicity in participation and academic attainment in UK higher education: An update. *Oxford Review of Education*, *46*(3), 346–362. https://doi.org/10.1080/03054985.2019.1702012

Sabnis, S., Yu, R., & Kizilcec, R. F. (2022). Large-scale student data reveal sociodemographic gaps in procrastination behavior. In *Proceedings of the Ninth ACM Conference on Learning at Scale* (L@S 2022), 1–3 June 2022, New York, New York, USA (pp. 133–141). ACM. https://doi.org/10.1145/3491140.3528285

Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(4), e1249. https://doi.org/10.1002/widm.1249

Sboev, A., Gudovskikh, D., Rybka, R., & Moloshnikov, I. (2015). A quantitative method of text emotiveness evaluation on base of the psycholinguistic markers founded on morphological features. *Procedia Computer Science*, *66*, 307–316. https://doi.org/10.1016/j.procs.2015.11.036

Sedrakyan, G., Malmberg, J., Verbert, K., Järvelä, S., & Kirschner, P. A. (2020). Linking learning behavior analytics and learning science concepts: Designing a learning analytics dashboard for feedback to support learning regulation. *Computers in Human Behavior*, *107*, 105512. https://doi.org/10.1016/j.chb.2018.05.004

Shahjahan, R. A., Estera, A. L., Surla, K. L., & Edwards, K. T. (2022). "Decolonizing" curriculum and pedagogy: A comparative review across disciplines and global higher education contexts. *Review of Educational Research*, *92*(1), 73–113. https://doi.org/10.3102/00346543211042423

Singhal, Y., Jain, A., Batra, S., Varshney, Y., & Rathi, M. (2018). Review of bagging and boosting classification performance on unbalanced binary classification. In A. Goswami (Ed.), *Proceedings of the 2018 IEEE Eighth International Advance Computing Conference* (IACC 2018), 14–15 December 2018, Delhi, India (pp. 338–343). IEEE. https://doi.org/10.1109/IADCC.2018.8692138

Skopec, M., Fyfe, M., Issa, H., Ippolito, K., Anderson, M., & Harris, M. (2021). Decolonization in a higher education STEM institution—is "epistemic fragility" a barrier? *London Review of Education*, *19*(1), 1–21. https://doi.org/10.14324/LRE.19.1.18

Sloan-Lynch, J., & Morse, R. (2024). Equity-forward learning analytics: Designing a dashboard to support marginalized student success. In *Proceedings of the 14th International Conference on Learning Analytics and Knowledge* (LAK 2024), 18–22 March 2024, Tokyo, Japan (pp. 1–11). ACM. https://doi.org/10.1145/3636555.3636844

Tajfel, H. (1970). Experiments in intergroup discrimination. *Scientific American*, *223*(5), 96–102. https://doi.org/10.1038/scientificamerican1170-96

Tajfel, H., Billig, M. G., Bundy, R. P., & Flament, C. (1971). Social categorization and intergroup behaviour. *European Journal of Social Psychology*, *1*(2), 149–178. https://doi.org/10.1002/ejsp.2420010202

Tajfel, H., Turner, J. C., Austin, W. G., & Worchel, S. (2000). An integrative theory of intergroup conflict. In M. J. Hatch & M. Schultz (Eds.), *Organizational identity: A reader* (pp. 56–65). Oxford University Press. https://doi.org/10.1093/oso/9780199269464.003.0005

Tate, T., & Warschauer, M. (2022). Equity in online learning. *Educational Psychologist*, *57*(3), 192–206. https://doi.org/10.1080/00461520.2022.2062597

The Pandas development team. (2020). Pandas-dev/pandas: Pandas. https://github.com/pandas-dev/pandas

Tincher, M. M., Lebois, L. A. M., & Barsalou, L. W. (2016). Mindful attention reduces linguistic intergroup bias. *Mindfulness*, *7*(2), 349–360. https://doi.org/10.1007/s12671-015-0450-3

Yu, R., Lee, H., & Kizilcec, R. F. (2021). Should college dropout prediction models include protected attributes? In *Proceedings of the Eighth ACM Conference on Learning at Scale* (L@S 2021), 22–25 June 2021, online (pp. 91–100). ACM. https://doi.org/10.1145/3430895.3460139