Towards an automatic approach for uncovering ethnic bias in online learning texts

Josmario Albuquerque

A thesis submitted to The Open University for the degree of

Doctor of Philosophy

Institute of Educational Technology
Faculty of Wellbeing, Education and Language Studies
The Open University

March 2023
Abstract

Recent findings have indicated persistent ethnic biases in online learning platforms. For instance, researchers have suggested that individuals from ethnic minority groups are more likely to be marginalised and have their academic performance diminished. Although compelling evidence indicates the impact of ethnic biases on students, uncovering such biases in online learning platforms is challenging. First, the large amounts of online educational data make it impractical to uncover such biases by hand. Second, bias can be subjective which means that what is considered biased in certain contexts might not be considered biased in others. In addition, individuals from certain cultures and groups might perceive bias differently.

Accordingly, this PhD thesis aims to answer the following overreaching research question: How might ethnic bias in text-based online learning materials be automatically identified while considering the subjective nature of bias? A design-based research methodology is adopted to answer that question, where several cycles of design, implementation, and reflection are performed. This iterative process is organised into three studies.

In Study 1, literature for existing approaches aiming at uncovering bias in texts is reviewed. Promising computer-based approaches to identify bias in texts are selected based on the literature review and implemented in the context of an online learning platform. Drawn from the limitations of those approaches, Study 2 looked at how ethnic biases manifest in learning texts by asking 193 higher education students to label ethnic bias in Open Educational Resources (OER), and how contextual elements like the OER’s title and discipline help them identify such biases. In Study 3, well-known learning analytics models are applied to the labelled dataset from Study 2, and their performance is checked against the identification of perceived ethnic bias in textual OERs.

Key findings from Study 1 reveal bias in online learning texts has received limited attention from the research community, in particular, selected studies delineate bias based on
its theoretical aspects rather than how students perceive it. Study 2 suggests that students from ethnic minority populations perceive ethnic bias differently than students identified as White, as a range of sentences from selected online learning texts is labelled as biased by one group but not by the other. Study 3’s key findings indicate statistically significant correlations between perceived ethnic bias and socio-linguistic features like socialisation and aggressiveness. Those features are used for training different classifiers against the identification of ethnic bias. The results suggest SVMs and Random Forest models reliable in identifying ethnic bias in online learning texts. Combining logistic regression, SVM, Naive Bayes, K-nearest Neighbors, and XGBoost could also provide balanced performance. Naive Bayes may not be as effective as SVMs and Random Forests in identifying bias but had the best precision when dealing with unknown data.

Overall, this PhD thesis has made a substantial contribution to understanding and identifying potential ethnic bias in online learning texts, which is crucial for reducing inequities and promoting inclusiveness in online learning platforms. Furthermore, this thesis has contributed to knowledge in Learning Analytics by providing evidence of how ethnic bias manifests in online learning texts and how certain computational models might support the automatic identification of such biases in large datasets.

**Keywords**— Learning analytics; ethnic bias; online learning; bias in text
To my parents.

Thank you for your constant support.
Acknowledgements

First, I would like to express my sincere gratitude to each of my supervisors, Prof. Bart Rienties, Dr Wayne Holmes, and Dr Martin Hlosta, for their incredible support, guidance, and inspiration throughout this project.

Also, I would like to thank those who examined this work during the upgrade and at its final stage. The feedback received was extremely valuable.

Thanks to The Open University for funding my studies and allowing me to be a part of its amazing community. In particular, thanks to the admin personnel at the IET, WELS faculty and Research Degrees for their kind and prompt support when needed. Also, thanks to all my peers at the IET and WELS for their company, laughs, chats, and encouragement.

Thanks to all of you who, directly or indirectly, contributed to this project.

Finally, my profound gratitude to my family and friends for their unconditional support and love.
Declaration of Authorship

I hereby declare that I am the sole author and composer of this thesis and that it has not been submitted, either in whole or in part, in any previous application for a degree. Except where otherwise acknowledged, the work presented is entirely my own.

References to relevant work


• **Albuquerque, J.,** Rienties, B., Holmes, W., & Hlosta, M. (2023). *What is biased, and according to whom? Students’ perspectives of ethnic bias in open educational resources* (In preparation)

# Table of Contents

## 1 Introduction

1.1 Context ....................................................... 2
1.2 Problem ................................................... 6
1.3 Objectives .................................................... 9
1.4 Thesis structure .......................................... 12
1.5 Chapter summary ......................................... 14

## 2 Background

2.1 Chapter overview ........................................... 15
2.2 Rationale ................................................... 15
2.3 Online learning ............................................ 18
    2.3.1 Fundamentals ........................................ 18
    2.3.2 Challenges ........................................... 20
2.4 Bias fundamentals ........................................ 21
    2.4.1 Social identity & group categorisation .............. 23
    2.4.2 Bias as prejudice, discrimination, and stereotype .... 25
    2.4.3 Subjective aspects of bias ......................... 27
2.5 Bias in online learning ................................... 29
    2.5.1 Sources of bias ..................................... 29
    2.5.2 Implications for learning ............................ 33
2.6 Bias identification ........................................ 35
    2.6.1 Non-automated strategies .......................... 35
    2.6.2 Automated strategies ............................... 36
    2.6.3 Limitations of existing strategies ................. 39
2.7 Learning analytics ........................................ 41
    2.7.1 LA and complex issues in online learning .......... 42
2.8 Chapter summary ......................................... 44
3 Methodology

3.1 Chapter overview ........................................... 45
3.2 Research philosophy ........................................... 45
   3.2.1 Ontology, Epistemology, and Axiology ................. 46
   3.2.2 The present project within a pragmatic paradigm .... 48
   3.2.3 Methodological choice .................................... 51
3.3 Underlying methodology ....................................... 51
   3.3.1 Fundamentals of DBR ....................................... 52
   3.3.2 DBR for educational technologies .......................... 55
   3.3.3 Challenges and criticisms of DBR .......................... 57
   3.3.4 DBR and this research ...................................... 59
3.4 Research design ............................................... 60
   3.4.1 Key research methods ...................................... 61
   3.4.2 Study 1 ..................................................... 63
   3.4.3 Study 2 ..................................................... 65
   3.4.4 Study 3 ..................................................... 70
3.5 Data & Ethics ................................................. 74
   3.5.1 Ethics review ............................................... 74
   3.5.2 Data protection ............................................. 74
3.6 Chapter summary ............................................... 76

4 Study 1 ........................................................................ 77
4.1 Chapter overview ............................................... 77
4.2 Research Cycle 1 .................................................. 77
   4.2.1 RC1: Analysis and Exploration ......................... 77
   4.2.2 RC1: Design and Construction ......................... 79
   4.2.3 RC1: Evaluation and Reflection ......................... 91
4.3 Chapter summary ............................................... 101

5 Study 2 .................................................................. 102
5.1 Chapter overview ............................................. 102
5.2 Research Cycle 2 ............................................. 103
   5.2.1 RC2: Analysis and Exploration ....................... 103
   5.2.2 RC2: Design and Construction ....................... 105
   5.2.3 RC2: Evaluation and Reflection ..................... 116
5.3 Research Cycle 3 ............................................. 119
   5.3.1 RC3: Analysis and Exploration ....................... 120
   5.3.2 RC3: Design and Construction ....................... 121
   5.3.3 RC3: Evaluation and Reflection ..................... 124
5.4 Chapter summary ............................................. 133

6 Study 3 ....................................................... 134
   6.1 Chapter overview ......................................... 134
   6.2 Research Cycle 4 ......................................... 135
      6.2.1 RC4: Analysis and Exploration ..................... 135
      6.2.2 RC4: Design and Construction ..................... 141
      6.2.3 RC4: Evaluation and Reflection ..................... 146
   6.3 Research Cycle 5 ......................................... 151
      6.3.1 RC5: Analysis and Exploration ..................... 151
      6.3.2 RC5: Design and Construction ..................... 153
      6.3.3 RC5: Evaluation and Reflection ..................... 158
   6.4 Chapter summary ......................................... 163

7 Discussion ..................................................... 165
   7.1 Chapter overview ......................................... 165
   7.2 Key findings .............................................. 166
      7.2.1 Study 1 ............................................. 166
      7.2.2 Study 2 ............................................. 168
      7.2.3 Study 3 ............................................. 170
   7.3 Limitations ............................................... 172
# List of Figures

2.1 Group interactions .................................................. 22
2.2 Inter-group bias and related concepts .............................. 27
2.3 Sources of bias ....................................................... 30

3.1 Processes of DBR ..................................................... 54
3.2 Research cycles ....................................................... 61

4.1 Overview of RC1 ...................................................... 80
4.2 Example of course transcription ..................................... 81

5.1 Instructions to participants .......................................... 110
5.2 Sample question ....................................................... 111
5.3 Attention check question ............................................ 113
5.4 Ethnic bias across groups ........................................... 125
5.5 Thematic analysis ..................................................... 130

6.1 Example of language abstraction ................................... 137
6.2 Overview of feature expansion ..................................... 145
List of Tables

2.1 Example of gender bias ........................................... 37
2.2 Example of subjective bias ...................................... 38
3.1 Research paradigms .............................................. 48
3.2 Overview of the research methods .............................. 61
3.3 Overview of Study 1 .............................................. 65
3.4 Overview of Study 2 .............................................. 69
3.5 Overview of Study 3 .............................................. 73
4.1 Sentences assessed for potential biases ....................... 90
4.2 Sentences identified as ‘biased’ ................................. 92
4.3 Possibility of random classification ........................... 94
4.4 Number of marks each sentence received .................... 95
4.5 Approaches vs Participants .................................... 96
4.6 Approaches’ accuracy based on participants’ answers ....... 97
5.1 Keywords referring to ethnic groups ......................... 106
5.2 Participants’ selection criteria .................................. 108
5.3 Goals, questions, and metrics that guided RC2. ......... 112
5.4 Examples and descriptions of contextual attributes. ...... 117
5.5 Metrics computed in RC2 ........................................ 118
5.6 Sentences marked as biased ..................................... 126
5.7 Potential reasons for ethnic bias ............................... 128
6.1 Psycholinguistic Features ....................................... 137
6.2 Linguistic abstraction and behaviour description .......... 138
6.3 Linguistic Abstraction Features ................................. 139
6.4 Features about Group Mentions and Valence ............. 142
6.5 Context Features .................................................. 142
6.6 Feature selection: correlation analysis ........................................ 148
6.7 Tested hyper-parameters ......................................................... 157
6.8 Hyper-parameters after fine-tuning ........................................... 158
6.9 Model’s performance after fine-tuning ....................................... 159
6.10 Models’ performance on unknown data .................................... 160
6.11 Model recommendations for bias detection in texts .................... 163
List of Abbreviations

AMT ........ Amazon Mechanical Turk
ANOVA ........ Analysis of Variance
AWS ........ Amazon Web Services
BERT ........ Bidirectional Encoder Representations from Transformers
BME ........ Black and Minority Ethnic
DBR ........ Design-Based Research
GDPR ........ General Data Protection Regulation
GQM ........ Goal-Question-Metric
HREC ........ Human Research Ethics Committee
IAT ........ Implicit Association Test
ICT ........ Information and Communication Technology
KNN ........ K-Nearest Neighbours
LA ........ Learning Analytics
ML ........ Machine Learning
MOOCs ........ Massive Open Online Courses
NLP ........ Natural Language Processing
NLTK ........ Natural Language Toolkit
NPOV ........ Neutral Point of View
OER ........ Open Educational Resource
OU ........ The Open University
PCA  ..........  Principal Component Analysis
POS  ..........  Part-of-Speech
SIT  ..........  Social Identity Theory
SVMs  ..........  Support vector machines
VADER  ........  Valence Aware Dictionary and sEntiment Reasoner
VLE  ..........  Virtual Learning Environment
Illustrative scenario:

Two students, Student 1 who is African British and Student 2 who is White British, are reading the same homework assignment about the U.S. Independence for their online class on American History. The homework includes the following sentence:

“About twelve years ago I hired a whaleboat and four Black men and proceeded to Long Island after a load of round clams.”

Student 1, as an African British, might perceive the sentence as a racist remark against the African British community arguing that there was no need for the author to specify the man’s ethnic group. In contrast, Student 2 might interpret the sentence as a historical fact and feel that the sentence is appropriate for the homework assignment given its historical context.

Despite this scenario being fictional, the sentence above was extracted from an actual online learning material and was considered inappropriate by certain ethnic groups and appropriate by others. This scenario also illustrates how individuals from distinct ethnic groups might perceive bias differently in the same text, and that it is important for authors of learning materials (e.g., course designers) to be mindful of how their content might be interpreted by different ethnic groups.

Illustrative scenario (continuation):

While Student 1 and Student 2 perceived the statement differently, both agree that learning materials should not discriminate or be biased against anyone. Therefore, they decide to talk to their tutors.

The tutors feel apologetic for knowing their students felt discriminated against, especially when that was not their intention (i.e., by bringing such ethnic content to the class, they expected to raise student awareness about ethnic issues). The tutors look at the material more closely and realise it was taken from a database of open edu-
cational resources — a database created with the aim of making learning resources more accessible to all and containing more than 50 thousand learning materials!

They then realise that reviewing that massive number of learning materials for potential biases would demand a huge effort. In addition, they are not sure how to decide which materials should be adjusted as what they think might be appropriate may be considered biased by others (and vice versa).

This illustrates how certain biases might be incorporated into texts even when the author’s intention is not to discriminate. It also highlights the challenge of addressing potential biases within online learning materials given the massive amount of data and the fact that different groups might perceive biases differently.

Accordingly, the effort put into this PhD project was directed towards a way to mitigate such issues. In particular, this PhD thesis focused on the possibility of automating the identification of those biases in online learning texts while accounting for the perspectives of different individuals. The next section (Section 1.1) provides more details about the present research within the broader context. Then, Section 1.2 details the research problem and how it led to the aims of this doctoral project.

1.1 This thesis within the broader context

The number of students taking online classes has substantially risen in the last decades and has drawn the attention of many researchers (Allen & Seaman, 2007; College, 2022). To illustrate, a recent report suggests there are more than 100 million students enrolled in online courses worldwide (College, 2022). In addition, the ongoing COVID-19 pandemic (Organization, 2019) has also contributed to the increased number of students attending their classes remotely, where many of them who used to attend their classes in person were “forced” to engage in online activities (Beaunoyer et al., 2020).

For a large number of students on online platforms, such settings impact them in
different ways. For example, researchers argued that online learning can benefit
students by improving their interactions through instant responses to their ques-
tions (X. Chen et al., 2018; Elmahdi et al., 2018; Zhu et al., 2020), by providing
a variety of resources to make learners more engaged (Buckley & Doyle, 2016;
Cohen et al., 2018; Johnston et al., 2018; Looyestyn et al., 2017), and/or by giving
teachers the opportunity to inform their decisions based on real-time data about their
students (Herodotou et al., 2019; Kuzilek et al., 2015; Papamitsiou & Economides,
2016; Rienties & Jones, 2020).

In contrast to these possible positive affordances of online learning, there are also
several challenges, such as high dropout rates (Grau-Valldosera et al., 2019; Xavier
& Meneses, 2020), students’ lack of motivation (Alonso-Mencía et al., 2020; Lee
et al., 2019), and prevailing educational inequities experienced by students across
various demographic groups, including but not limited to gender, social class, and
ethnicity (Goudeau et al., 2021; Richardson et al., 2020; Welser et al., 2019). Over-
all, despite the possible benefits of online learning for students, those ongoing
challenges can diminish some of the benefits and impair student experiences and
achievements.

Looking at ways to improve students’ experiences in those platforms, researchers
proposed various means to mitigate some of those issues, particularly by employ-
ing educational technologies (Bovermann & Bastiaens, 2020; Hlosta et al., 2021;
Namoun & Alshanqiti, 2020; Q. Nguyen et al., 2020; Sedrakyan et al., 2020). For
example, a recent systematic literature review suggested that dropouts can be pre-
vented using Learning Analytics (LA) (de Oliveira et al., 2021), which is “the mea-
surement, collection, analysis and reporting of data about learners and their con-
texts, for purposes of understanding and optimising learning and the environments
in which it occurs” (Long et al., 2011, p. 3). Other studies used features from games
(e.g., points, badges, rankings) to keep students motivated and engaged (Bovermann
& Bastiaens, 2020; Ertan & Kocadere, 2022).
Nonetheless, despite the huge effort already made to improve online learning platforms, recent findings suggested that ongoing inequities in those settings need special attention from the research community (Bayrakdar & Guveli, 2020; Q. Nguyen et al., 2020; Sabnis et al., 2022). To illustrate, Q. Nguyen et al. (2020) analysed data from nearly 150k students across 401 online courses and identified a discrepancy in academic performance between White students and those from ethnic minority groups (i.e., those ethnic groups with the smallest percentages of students\(^1\)). These findings highlight the importance for researchers and educational practitioners to continue exploring ways to address the persistent inequities in online learning and ensure that all students have equal access to quality education.

Inequities in online learning can be caused by a range of factors, such as limited access to technology (Beaunoyer et al., 2020; Ferri et al., 2020), disparities in socioeconomic status (Francis & Weller, 2022), and social biases (Ashong & Commander, 2012; Jacoby-Senghor et al., 2016; Pennington et al., 2016). For clarity, while “bias” has different meanings across a variety of contexts, in this thesis, “bias” refers to “inter-group bias”, which is an inclination against (negative bias) or in favour of (positive bias) individuals of particular groups based on their social characteristics (Greenwald et al., 1998; Taylor & Doria, 1981), such as gender (gender bias) or ethnic group (ethnic bias).

Accordingly, this doctoral project was motivated by inequities in online learning caused by negative inter-group biases. This motivation emerged from the systemic and pervasive nature of bias (Sabnis et al., 2022). For example, while factors such as limited access to technology and socioeconomic disparities were widely recognised as important contributors to educational inequities, bias can affect students regardless of their socioeconomic status or access to resources and can have lasting effects on their academic achievement and personal development (Doubé & Lang, 2012; 2022).

\(^1\)To benefit clarity throughout this thesis, the term “minority” will be used in its ordinary meaning (i.e., in the numerical sense). Accordingly, “ethnic minority” refers to ethnic groups that, according to the UK 2021 Census and the US 2020 Census, are numerically smaller than White British (in the UK) and White Americans (in the US).
Jacoby-Senghor et al., 2016; Jordano & Touron, 2017; H. Nguyen & Ryan, 2008; Pennington et al., 2016). To illustrate, Pennington et al. (2016) conducted a systematic literature review and found 45 experiments from 38 articles where bias against different groups (e.g., females, African-Americans) triggered several psychological mechanisms (e.g., anxiety, negative thinking, cognitive load) that affected student academic performance.

Beyond affecting students directly, researchers also showed that biases can be incorporated in curricula (Shahjahan et al., 2022; Skopec et al., 2021), and in learning technologies (R. S. Baker & Hawn, 2021; Deho et al., 2022). To illustrate, Shahjahan et al. (2022) analysed 207 articles related to bias in curricula and showed that, when developing new learning materials, particular cultural perspectives were introduced in the learning materials. In other words, the authors showed the curriculum was being framed in a way to represent the views of White individuals to the detriment of the perspectives of stigmatised social groups (e.g., Black females). R. S. Baker and Hawn (2021) reviewed studies about bias in online learning and highlighted the presence of bias in different aspects of learning platforms, such as in predictive models trying to forecast student dropouts and failing rates, in automated essay scoring, and assessments of language proficiency.

Accordingly, given the various ways in which inter-group bias can be incorporated into online learning, a notable delimitation at the beginning of this PhD project was focusing on bias in text-based learning materials (e.g., textbooks, lesson plans, video transcriptions, etc.). This decision is justified by those materials being a primary source of content in various educational settings, both online and traditional (Mayer, 2019). In addition, the literature indicates that biases in text-based materials can lead to negative learning outcomes and perpetuate stereotypes, thus affecting the academic performance and personal development of students from various backgrounds (Shahjahan et al., 2022; Skopec et al., 2021).

Overall, despite those findings showing the consequences of bias for students, iden-
CHAPTER 1. INTRODUCTION

1.2 Problem Definition

Researchers already proposed various means to identify bias, including questionnaires, discourse analysis, and natural language processing. For example, Falchikov and Magin (1997) checked the presence of gender bias in students’ assessments using a questionnaire in which students were grouped according to their gender and asked to mark other students’ assessments. Their findings indicated a small effect favouring females, which means that respondents gave females higher grades. While their findings provided insightful evidence regarding bias in educational settings, the questionnaire was not validated, which weakens the reliability of the results.

With the aim of providing a more reliable solution to assess bias, a recent study (Gaasedelen et al., 2019) focused on developing a cognitive bias scale based on a Personality Assessment Inventory. After sampling more than 300 individuals, the researchers applied Item Response Theory to refine the scale. The findings suggested the scale could be useful for assessing bias in a healthcare setting, however, this might not be adequate for educational contexts. In addition, the use of questionnaires may not be effective for measuring subjective constructs such as inter-group bias, given that they rely on respondents’ answers, which are subject to self-report bias (Althubaiti, 2016).
In contrast, P. Baker (2012) and Qayyum et al. (2019) looked for potential inter-group bias in texts by employing corpus linguistics methods, which comprise the analysis of large amounts of text for studying language (Kennedy, 2001). For example, P. Baker (2012) investigated the British Press regarding extremist statements towards Islam and/or Muslims. The methodology comprised analysing major newspapers (e.g., The Guardian), and they found that Muslims were commonly associated with concepts of extreme belief rather than moderate belief. Recently, other researchers (Qayyum et al., 2019) also considered discourse analysis to investigate biases, but in the context of social media. In that research, they analysed sentiments, toxicity, and biases in Tweets related to the Pakistani and Indian political parties. Their findings identified evidence of media bias and discourse toxicity in the political discourse about those nations. Despite these findings, neither approach was validated in educational settings, which raises questions about their suitability for use in such contexts. In addition, while those methods seem more reliable for identifying bias in language when compared to self-reports, the implementation of such approaches requires a huge human effort to be applied to large texts since they were not created to be executed automatically.

In another study conducted in the context of Wikipedia (Recasens et al., 2013), researchers reinforced the assumption that manually identifying bias in texts is complex and stressed the importance of automating the process. They combined discourse analysis with statistical methods to identify framing and epistemological biases – certain types of bias related to perspective-specific words and presupposed truths in text (Recasens et al., 2013). In the end, they provided a set of language characteristics (e.g., positive and negative words) that were used to indicate which parts of Wikipedia’s articles were potentially biased. Subsequently, such characteristics were used by Pryzant et al. (2020) to identify and neutralise biased language using supervised learning – a computational approach in which algorithms are trained to predict output values based on a set of input data (Russell, 2010).
Pryzant et al. (2020) focused on subjective biases in Wikipedia, which they defined as a type of bias that creates inappropriate subjectivity in language, introduces opinion, feeling, or taste, and makes texts potentially unfair and skewed. Pryzant et al. (2020) also considered the Bidirectional Encoder Representations from Transformers (BERT) model – a natural language processing technique proposed by Devlin et al. (2018) – to identify and edit problematic words in each corpus. Their findings indicated the approach may be suitable to Wikipedia and news, and they also showed evidence supporting the predominance of biases in areas like history, politics, philosophy, sports, and language.

Although the algorithm produced valuable results, it only addresses bias at the word level, leaving biases in paragraphs and entire sentences unaddressed. Furthermore, neither approach has been validated in educational settings, so their effectiveness in such contexts remains unknown. It is also worth noting that subjective bias may not necessarily pose a problem in educational texts. For instance, while statements expressing opinions may be considered subjectively biased and thus inappropriate for encyclopaedias, they could be acceptable in educational contexts. This notion aligns with Tajfel’s research on group categorisation and discrimination, which demonstrates that individuals can categorise themselves based on various attributes, leading to complex and fluid group boundaries that may require understanding specific contexts to define them (Tajfel, 1970; Tajfel et al., 1971; Turner et al., 1994). In other words, what constitutes bias in learning texts may depend on the perspectives of specific groups within contexts, as illustrated in the introductory scenario.

Overall, substantial efforts were made towards developing an appropriate way to identify bias in text, both manually and automatically, however, the proposed approaches have several limitations. As indicated, manual strategies are not practical for a huge amount of data such as in online learning environments because it would require an unfeasible human effort to analyse text by hand. Meanwhile, the automatic approach proposed by Pryzant et al. (2020) only focused on subjective
biases, so it may not be entirely suitable for identifying bias against specific groups of individuals (i.e., inter-group bias). Beyond that, their approach only addresses subjective bias at a word level, which also makes it limited when looking at full sentences or paragraphs.

Accordingly, to the best of my knowledge, there is no effective approach to automate the identification of inter-group bias in online learning texts while accounting for the context of those texts and the perspectives of particular groups. These limitations informed the research conducted during this doctoral project. In particular, the thesis focused on the possibility of automating the identification of ethnic bias in online learning texts while accounting for subjective aspects of bias like context and individuals’ perspectives. Although other types of bias have their importance in online learning, ethnic bias was chosen for being a widespread issue that has been shown to impact students’ access to opportunities, resources, and social relationships (Ashong & Commander, 2012; Q. Nguyen et al., 2020; Richardson et al., 2020; Tate & Warschauer, 2022).

Below, more details regarding the objectives of this project are presented as well as how each objective is aligned with each research question.

1.3 Objectives & Research Questions

As indicated above, this PhD thesis aims to support the identification of ethnic bias in online text-based learning materials considering contextual aspects of ethnic bias, and the perspectives of students from different ethnic groups. In other words, this thesis focused on answering the overreaching research question:

How might ethnic bias in text-based online learning materials be automatically identified while considering the subjective nature of bias?

By “subjective nature of bias” I mean accounting for contextual aspects of learning materials and the perspectives of students from different ethnic groups. Accord-
ingly, to answer that overreaching question, three main objectives were defined, leading to three main research questions:

1. To understand to what extent existing approaches aimed at bias identification in text from non-educational settings perform in online learning materials. This objective comprised identifying existing computational approaches focusing on bias identification and then implementing them in educational settings so that further limitations present in such approaches could be identified. This objective led to the first research question:

   **RQ1.0:** How do computational approaches, that claim to uncover inter-group bias in non-educational texts, perform in text-based online learning materials?

2. To understand how ethnic bias manifests in online text-based learning materials according to the perspectives of students from a variety of ethnic groups, and to what extent contextual elements like the material’s title and discipline help them identify such biases. This objective was motivated by existing evidence suggesting that what is considered “biased” might depend on the extent to which individuals identify themselves within a particular group (Tajfel et al., 1971), and given a social context (Hogg, 2016). Accordingly, this objective focused on answering the following research question:

   **RQ2.0:** How does bias against certain ethnic groups manifest in text-based online learning materials?

3. To implement and evaluate different computational approaches that could support the automation of bias identification in online text-based learning materials. Such an objective focused on: (i) identifying relevant attributes (also known as “features”) in text-based online learning materials that could be linked to inter-group bias; and (ii) testing supervised learning models that could assist in the classification of bias in those materials. In essence, this
CHAPTER 1. INTRODUCTION 1.3. OBJECTIVES

objective focused on answering the third research question:

RQ3.0: To what extent can Learning Analytics assist in the identification of ethnic bias in text-based online learning materials?

To achieve those objectives, this project adopted a design-based research (DBR) methodology (Baumgartner et al., 2003; Wang & Hannafin, 2005), which is known as a flexible methodology that comprises a series of iterative cycles of design, implementation, and evaluation (Wang & Hannafin, 2005). DBR was shown as a promising methodology to connect theory and practice in Education (Anderson & Shattuck, 2012), and it has been used extensively by researchers to design and improve learning settings (Cowling & Birt, 2018; Dolmans, 2019; Koivisto et al., 2018; Vanderhoven et al., 2016; Wolcott et al., 2019).

Therefore, this PhD project adopted DBR based on the need for a flexible methodology that could connect the theories of inter-group bias to a practical mechanism to support the identification of such biases in online learning materials. Accordingly, this thesis consists of three main studies (Study 1 – Study 3), each study focusing on one objective respectively.

While details about this methodological approach are provided in Chapter 3, the three studies can be summarised as follows.

Study 1 (Chapter 4) was completed in one research cycle (RC1) and focused on implementing existing approaches aiming at uncovering general inter-group bias in online texts. While the current PhD research focused on ethnic biases, the scope of Study 1 was expanded to a wider range of approaches (i.e., to include other inter-group biases) aiming at a more comprehensive understanding of how the identification of ethnic bias in text could be automated. Accordingly, two approaches were selected from the literature and implemented in the context of an online learning platform.

Study 2 (Chapter 5) was completed in two research cycles (RC2 and RC3) in which
students were recruited to assess ethnic bias in online text-based learning materials. They also provided feedback on the extent certain contextual elements of online learning materials (e.g., their title and discipline) help them identify such biases. Accordingly, Study 2 provided insights into how ethnic bias manifests in learning texts and provided a labelled dataset containing potential biases.

Drawn from the results of Study 2, Study 3 (Chapter 6) was also completed in two research cycles (RC4 and RC5) and focused on applying well-known Learning Analytics models to a dataset generated in Study 2 to see how they would perform in a bias-identification task. Finally, the performance of such models was checked against the potential biases identified by the students in Study 2.

In the next section, further details about the organisation of this thesis are presented.

1.4 Thesis structure

The remainder of this thesis is organised as follows:

• **Chapter 2: Theoretical Background**

  This chapter delves into the theoretical foundation that underpins the research presented in this study. It starts by presenting the rationale used to review the literature and the key assumptions that guided this PhD project. Then, Section 2.3 focuses on opportunities and challenges of online learning platforms, Section 2.4 presents the fundamentals of inter-group bias, which is followed by Section 2.5 highlighting the impact of such biases on student achievements. Section 2.6 covers existing approaches that are aimed at the identification of bias in text, and Section 2.7 focuses on Learning Analytics.

• **Chapter 3: Methodology**

  This chapter focuses on presenting the methodological approach for this PhD project and general aspects of the design of each study (i.e., details of the
procedures and the materials of each study are covered in Chapters 4–6). Accordingly, the initial sections cover the research philosophy (Section 3.2) and the underlying methodology (Section 3.3) that supported the operationalisation of the research questions. After discussing the research philosophy and underlying methodology, Section 3.4 presents the research design, i.e., it describes the core studies of this doctoral project. In the end, Chapter 3 highlights ethical and data protection issues related to the execution of this methodology (Section 3.5).

• **Chapter 4: Study 1**

This chapter presents the initial study conducted during this PhD project (i.e., Study 1). Study 1 was motivated by limited evidence regarding the automation of identifying inter-group bias in texts. Accordingly, Section 4.2.1 explores the problem addressed in Study 1. Section 4.2.2 presents the research design, which includes the steps undertaken to uncover existing computational approaches for bias identification, the methods used for implementing them, and the procedures followed to evaluate them in an educational context. Then, Section 4.2.3 presents the results, and discussion, and indicates core implications for Study 2.

• **Chapter 5: Study 2**

This chapter aims to address the limitations of the approaches highlighted in Study 1 (Chapter 4). It also focuses on understanding how ethnic biases in online learning materials are perceived by students of different ethnic groups. This chapter also aims to help understand how contextual elements could affect the identification of ethnic bias in online learning materials. In addition, the findings presented in this chapter include a dataset indicating how ethnic biases look in online text-based learning materials, which was used as a baseline for training Learning Analytics models in Study 3 of this research project (Chapter 6).
• **Chapter 6: Study 3**

This chapter is informed by the findings of Study 2 (i.e., the evidence suggesting how ethnic biases manifest in text-based learning materials). Accordingly, Chapter 6 comprises the use of Learning Analytics models to automatically identify potential ethnic bias in text-based learning materials. In addition, this chapter is divided into two main parts: (i) the first part (Section 6.2) focuses on selecting features that might be useful for sporting ethnic bias in texts; then, (ii) Section 6.3 comprises testing commonly used classification methods against the identification of ethnic bias in online text-based learning materials.

• **Chapter 7: Discussion & Conclusions**

This chapter covers a general discussion regarding the findings presented in the previous chapters. Section 7.2 discusses the key findings of this PhD project. Afterwards, Section 7.3 focuses on limitations. Then, Section 7.4 highlights theoretical, practical, and methodological implications. Finally, concluding remarks are presented in Section 7.5.

### 1.5 Chapter summary

This chapter covered an overview of the present thesis. Initially, the context for this PhD project was presented, which comprised an overview of online learning platforms and ongoing social issues in such settings. Then, the key research problem was defined by indicating the challenges of identifying inter-group bias in the text. Afterwards, the aims and objectives of this project were presented, which were followed by the respective research questions. Finally, the overall structure of this PhD thesis was outlined. Now that those preliminary aspects underlying this thesis were covered, the next chapter focuses on the theoretical background of this PhD project.
2 | Theoretical Background

2.1 Chapter overview

This chapter covers the theoretical background in which this research is grounded. The next section (Section 2.2) presents the overall rationale used to review the literature and summarises the research assumptions that guided this PhD project. Then, the remainder of this chapter is organised according to each research assumption. For instance, as mentioned in the previous chapter (Chapter 1), this PhD project focuses on negative ethnic bias against individuals of different ethnic groups. In particular, biases present in text-based learning materials that, according to students, can make individuals from those groups feel excluded or underrepresented. Accordingly, in order to establish a context for this project, Section 2.3 focuses on the opportunities and challenges of online learning platforms. Then, Section 2.4 deepens the theoretical background related to the research problem, i.e., it presents the fundamentals of group bias and its impact on student achievements. The following section, Section 2.6, covers existing approaches that are aimed at the identification of bias in text. Finally, Section 2.7 concentrates on Learning Analytics, which has been shown as a promising way to support equitable educational settings.

2.2 Literature review rationale

To identify relevant and high-quality studies for establishing a solid foundation for research, it is essential to review the literature systematically (Tranfield et al., 2003). This process enables researchers to identify patterns, trends, and gaps in the literature, ultimately contributing to a more comprehensive understanding of the research topic (Fink, 2014). Thus, aiming at establishing the theoretical background for this doctoral project, the following strategy, informed by the guidelines proposed by Booth et al. (2016), was adopted:
1. Define the research question: Each main topic covered in this chapter was guided by a specific question (see Appendix A.1).

2. Identify keywords: Keywords were identified based on the question identified in Step 1.

3. Construct search string: Search strings were constructed using the identified keywords and potential synonyms.

4. Perform the search: Searches were conducted on Google Scholar\(^1\), as it indexes studies from various databases.

5. Screen and select articles: Initially, titles and abstracts were checked against the research question (i.e., for relevance). Then, articles that were not in English or had no full-text availability were excluded (institutional access was attempted).

6. Assess article quality: To evaluate the quality of selected articles, the following factors were checked: the reputation of the journal or conference, the author’s expertise, the methodology used, and the relevance of the findings to the research question.

7. Extract and synthesise information: Key findings, such as the main benefits and challenges discussed in each article, were compiled.

8. Maintain a record of the search process: Although several records were stored in Mendeley\(^2\), the respective research questions and search strings can be found in Appendix A.1.

Once a systematic strategy for reviewing the literature was defined, the review process began. Accordingly, the literature review started in October 2019 by looking at the fundamentals of online learning along with its affordances and ongoing challenges, such as inequities and biases. That portion of the literature review led to the

\(^1\)https://scholar.google.com/
\(^2\)https://www.mendeley.com/
2.2. RATIONALE

following assumption:

Assumption 1: Despite the evidence informing the benefits of online learning platforms, ongoing social issues like inequities and biases make such settings less inclusive and unfair for certain students, in particular those from ethnic minority backgrounds.

Afterwards, based on that assumption, key theories which try to explain inequities and biases were reviewed. That portion of the literature review focused on the dynamics of social groups, and how biases in form of prejudice and stereotypes emerge in society. That review led to the next assumption:

Assumption 2: Bias is a complex issue that emerges from individuals’ perceptions regarding their social identity within a social group. Therefore, what is considered biased for a certain group might depend on their social contexts and cultures.

Once the key aspects of online learning and bias were reviewed, particular evidence regarding the impact of bias on students and their learning was sought. That portion of the literature review led to the third assumption:

Assumption 3: Inter-group bias\(^3\) is also present in online learning and, when not addressed, it can negatively affect students in different ways, for example, by diminishing their academic performance, and by misleading key decisions when such decisions are based on biased learning materials.

The next step in the literature review process focused on identifying what has been done to address biases in online learning platforms. Potential mechanisms to identify bias in such settings (automated and non-automated approaches) were searched for. That step led to the following assumption:

Assumption 4: Despite existing efforts to understand and avoid biases in non-

\(^3\)As indicated in the previous chapter (Section 1.1), inter-group bias refers to an inclination against individuals of particular groups based on their social characteristics (e.g., gender, ethnic origin). See Section 2.4 for more details.
educational contexts, mechanisms to address ethnic biases in online learning materials and mechanisms to address bias within large datasets are limited.

Finally, the literature was searched to understand how Learning Analytics has been used to address complex issues in online learning settings, thus highlighting its relevance for addressing issues like ethnic bias in online learning texts. This portion of the literature review led to the final assumption that grounds this project:

**Assumption 5: Learning Analytics seem promising for addressing subjective and complex issues in online learning.**

The next section provides more details about how the current literature informs each assumption.

### 2.3 Online learning

**Assumption 1: Despite the evidence informing the benefits of online learning platforms, ongoing social issues like inequities and biases make such settings less inclusive and unfair for certain students, in particular to those from ethnic minority backgrounds.**

This section focuses initially on the fundamentals of online learning. Then, some of the challenges found in such settings are covered.

#### 2.3.1 Fundamentals

Online learning is a broad term encompassing various concepts related to the use of technology in education. For example, distance learning and remote learning have been used to emphasise the fact students attend classes/courses remotely (i.e., from various geographic locations). Another widely used concept is Virtual Learning Environment (VLE), which was defined by Dillenbourg et al. (2002) as a virtual space for educational and social interactions in which students and tutors partici-
pate actively. With the rise of novel information and communication technologies (ICTs), that definition became widely used to designate Web-based VLEs (Tikam, 2013). Accordingly, VLEs can incorporate a range of technologies and resources, from text-based functions to video and augmented reality features, as highlighted by Keller (2005). In this thesis, the term “online learning” refers to general learning occurring via the World Wide Web (WWW), where students typically participate remotely.

Numerous studies have highlighted the positive impact of online learning platforms on student performance and achievement (Agudo-Peregrina et al., 2014; Arnold & Pistilli, 2012; Lu et al., 2018; Yu & Jo, 2014). For example, Yu and Jo (2014) employed a multivariate regression approach to pinpoint the factors that significantly influence the academic performance of undergraduate students. They found that attributes like time studying, collaboration, studying interval, and download of learning materials were statistically significant predictors of student success.

In a recent study, Lu et al. (2018) used a principal component regression approach to estimate the final academic performance of students in blended learning. Their predictive model was able to predict student success in the first one-third of the course duration, which could be used as an early intervention to prevent potential students from failure. Meanwhile, Agudo-Peregrina et al. (2014) used interaction data to predict student success in VLEs, supported face-to-face, and online learning. They found a statistically significant relationship between interactions and academic performance in online courses, but the results were not statistically significant for the other modalities.

It is also argued that online learning can support other pedagogical practices. For example, through the possibility of immediate feedback, improvement of collaboration, content personalization, and improvement of student satisfaction. To illustrate, Kangas et al. (2017) used a mixed-method approach to investigate the relation between teachers’ motivation and student satisfaction in a game-based environment.
and found that teachers’ pedagogical and emotional engagement had a statistically significant relation to student satisfaction with the online environment. Another study (Liaw et al., 2007) surveyed 168 participants in an online system and found that such educational settings can improve student autonomy. In addition, Jubaedah et al. (2020) analysed the usability of an online evaluation tool and showed that immediate feedback can facilitate the assessment process, which supports previous findings regarding immediate feedback (e.g., Scheeler and Lee (2002)).

Despite such benefits of online learning platforms, researchers have also highlighted ongoing challenges, which are presented in the subsequent section.

### 2.3.2 Challenges

As with any other technology, online learning also raised concerns in the scientific community. For example, there is evidence that when students’ skills do not meet technical requirements, they tend to have poor engagement within the online environment (Boisselle, 2014). Those findings reinforce previous research that indicated the importance of developing digital literacy (Alexander et al., 2012). They also highlight the importance of tutors improving psychological and socio-cultural knowledge to maintain a balanced relationship with students. Beyond that, Richardson et al. (2020) indicated there were persistent social issues within online learning environments, for example, social inequities related to gender, ethnicity, and social class.

In addition, because of the COVID-19 pandemic, online learning platforms have seen a massively increased number of students (Beaunoyer et al., 2020; Onyema et al., 2020), which has led to a larger number of students being exposed to the potential drawbacks of such platforms (Lemay et al., 2021). To illustrate, it has been shown that unsolved social inequities and biases can harm student performance (Pennington et al., 2016). For instance, Q. Nguyen et al. (2020) showed a persistent gap in student performance in an online learning setting, i.e., they showed...
that individuals from an ethnic minority background (e.g., Black and African students) were at more risk of not completing an online course successfully.

Other studies have suggested inter-group bias in online learning platforms. For example, Dawkins et al. (2017) suggested potential gender bias in physics questions from online courses at The Open University (UK). In essence, they found that certain question structures (e.g., language and format used) could potentially benefit one gender over another. In particular, the study showed that questions involving the interpretation of certain diagrams (e.g., two-dimensional graphs) were likely to promote male bias. A possible interpretation of those results is that biases could result from questions utilising language that is typically found in male-dominated fields, which might disadvantage female students who are not familiar with that terminology.

Overall, these findings highlight the importance of addressing such inequities in online learning platforms and informed the decision of concentrating this PhD thesis on inter-group bias in online learning. However, before highlighting particular implications of bias for students, key fundamentals of biases are presented.

### 2.4 Bias fundamentals

**Assumption 2:** *Bias is a complex issue that emerges from individuals’ perceptions regarding their social identity within a social group. Therefore, what is considered biased for a certain group might depend on their social contexts and cultures.*

The term *bias* is a broad concept that can be found in several domains to refer to different things. While in Statistics, for example, it may represent the underestimation or overestimation of a parameter, in Social Psychology it is commonly used to describe a preference (or favouritism) against (or in favour of) individuals of a social group (Greenwald & Krieger, 2006; Taylor & Doria, 1981). This concept of bias as a preference became popular with the investigations of Henri Tajfel about inter-
group categorisation and discrimination (Tajfel, 1970; Tajfel et al., 1971), when he found that individuals may categorise themselves based on a variety of attributes (e.g., gender, ethnic group, religion, or even trivial characteristics like a preference for paintings). He demonstrated that, based on those categorisations, people tend to favour members of their group (in-group) rather than members from other groups (out-group). Accordingly, bias can be positive (towards in-group members) or negative (against out-group members).

To illustrate, Figure 2.1 depicts in-group and out-group interactions as suggested by Fu et al. (2012). In essence, in-group interactions refer to interactions between members of the same social group (e.g., between members of Group B as indicated by red lines in Figure 2.1). In contrast, out-group interactions comprise the interactions between members of different social groups (e.g., between members of Group A and Group B as illustrated in Figure 2.1 in green colour).

As a result of social interactions and categorisations, inter-group biases can be present in a myriad of social settings, e.g., in court judgements (Gazal-Ayal & Sulitzceanu-Kenan, 2010), recruitment processes (Koval & Rosette, 2021), social media (Bliuc et al., 2018), healthcare (FitzGerald & Hurst, 2017), and educational settings (Mahmud, 2020; Quintana & Mahgoub, 2016; Van den Bergh et al., 2010; Warikoo et al., 2016), which is the context of this thesis.

While such biases can be present in a variety of contexts, this thesis focuses on
ethnic bias in text-based learning materials from online learning platforms. Furthermore, as also mentioned in Section 1.2, the present study concentrates on negative biases that can make individuals from particular ethnic groups feel excluded or discriminated against.

A key theory for understanding how inter-group bias is present in society and, therefore, in learning settings is the Social Identity Theory (SIT). Such a theory combines the findings of Tajfel’s studies along with other investigations conducted later. Hogg (2016) summarised SIT as:

\[ a \text{ unified conceptual framework that explicates group processes and inter-group relations in terms of the interaction of social cognitive, social interactive, and societal processes, and places self-conception at the core of the dynamic } \]

(p. 13).

As SIT combines vital concepts that are essential for understanding discrimination and biases in society, before diving into the implications of bias for online learning, critical aspects of SIT and their relation to group bias are presented.

### 2.4.1 Social identity & group categorisation

A key contribution of Tajfel’s research is the concept of "social identity" which is directly related to "social group". For instance, Tajfel et al. (1971) defined social identity as the knowledge of individuals that they belong to a social group and that membership has some emotional and value significance to them. That definition is also related to the concept of a social group, which can vary in size and provide their members with a shared social identity indicating who they are, what they should believe, and how they should behave (Hogg, 2016). Accordingly, an individual’s identity is based on their sense of belonging to a given social group.

Within this context, concepts such as “ethnic groups”, “ethnicity”, and “race” play a significant role in shaping individuals’ sense of belonging and shared social iden-
Ethnic groups are typically defined by shared cultural practices, ancestry, or other common characteristics (Barth, 1998), while ethnicity might refer to an individual’s affiliation with a particular ethnic group (Waters, 1990). Race, on the other hand, is often understood as a social construct that groups people based on physical characteristics (Omi & Winant, 2014; Smedley & Smedley, 2005). Accordingly, these classifications and affiliations contribute to the formation of social groups and the motivations behind individuals’ categorisation within these groups.

Researchers have tried to explain what motivates humans to categorise themselves in groups and different theoretical explanations have emerged in which three main motivations received special attention: optimal distinctiveness, self-esteem, and uncertainty. Leonardelli et al. (2010) suggested that group categorisation can be motivated by optimal distinctiveness, which is based on a balance between two main conflicting motives: (i) inclusion/sameness, satisfied by group membership; and (ii) distinctiveness/uniqueness, satisfied by individuality. In the self-esteem hypothesis (Abrams & Hogg, 2006; Tajfel et al., 1979), positive inter-group differentiation is known for elevating self-esteem, and low self-esteem is considered to motivate discrimination. Regarding uncertainty, Hogg et al. (2007) suggested that feeling uncertain about the world and how to behave is unsettling. In that case, group categorisation provides norms that describe how individuals behave and interact with each other, which is thought to reduce their uncertainty.

Those motives regarding group categorisation might also apply to ethnic groups, ethnicity, and race (Abrams & Hogg, 2006; Hogg et al., 2007; Leonardelli et al., 2010). To illustrate, Leonardelli et al. (2010) showed that individuals may seek to balance their need for inclusion within an ethnic group while maintaining their distinctiveness based on their unique cultural background. Similarly, ethnic or racial group membership may contribute to an individual’s self-esteem (Abrams & Hogg, 2006; Tajfel et al., 1979). In addition, Hogg et al. (2007) showed that self-categorisation based on ethnicity or race can help reduce uncertainty by providing
a shared set of norms and expectations.

Furthermore, Hogg (2016) suggested that social groups, including ethnic and racial groups, comprise categories that are mentally represented by individuals as “prototypes”, i.e., sets of interrelated attributes like attitudes, behaviours, customs, traditions, and physical characteristics that capture their overall similarities. Such prototypes are known to reinforce “the extent to which a group appears to be a distinct and clearly defined entity” (Hogg, 2016). While such group categorisation can provide self-assurance of one’s own identity within a group, it can also lead to several conflicts between individuals from the in-group and out-group. Those conflicts were also noted in Tajfel’s studies when investigating inter-group relations (Tajfel, 1970; Tajfel et al., 1971). These prototypes contribute to the distinctiveness and clarity of ethnic and racial group identities, further highlighting the importance of understanding the nuances of group categorisation in different contexts (Hogg, 2016).

Tajfel et al. (1979) argued that when individuals make comparisons between their own group and the out-group, they tend to distinguish their own group positively. In contrast, members of the out-group tend to be described as less favourable by in-group members, which is in line with Brewer and Campbell (1976) assumption in which inter-group relations are intrinsically in-group-favouring and ethnocentric. In addition, Hogg (2016) suggested that higher-status groups strive to protect their perceived superiority while lower-status groups fight to mitigate their social stigma and promote their positivity. Such a group dynamic, with different groups striving to protect their own (group) identity, can lead to biases in the form of prejudice, stereotypes, and discrimination. Those concepts are presented in the next section.

### 2.4.2 Bias as prejudice, discrimination, and stereotype

These four concepts (bias, prejudice, discrimination, and stereotype) are usually related and sometimes one is mistaken for the other. For instance, Wilder and Simon
(2001) suggested that inter-group bias had been viewed by other researchers as having three main components: prejudice, discrimination, and stereotype. Prejudice can be defined as “a hostile attitude or feeling toward a person solely because he or she belongs to a group to which one has assigned objectionable qualities” while discrimination is “acting on that negative prejudice” (Allport et al., 1954). Therefore, prejudice focuses on individuals’ attitudes while discrimination is particularly related to the action (or behaviour) resulting from such attitudes. In contrast, stereotype focuses on the categorisation of individuals into specific groups, which Hogg (2016) suggested as a type of “prototype”:

> If many people in one group share the prototype of their own or another group, the prototype is essentially a stereotype - if you alone believe that all Martians have skinny green bodies and huge heads, it’s a prototype, but if pretty much all other humans believe this, then the prototype is also a stereotype (p. 8).

Accordingly, as each dimension of inter-group bias is related to the way individuals perceive and categorise themselves, the essence of what is considered biased (or not biased) can depend on such categorisations. Therefore, the assessment of bias becomes subject to other factors like cultural and contextual aspects. Indeed, Hogg (2016) stressed the social context as a key element of the social identity theory: “Social identities also, very critically, highlight how the in-group is distinct from relevant out-groups in a particular social context” (p. 6).

Figure 2.2 suggests a more concise view of how those complex concepts aforementioned relate to each other. In essence, social identity explains how individuals categorise themselves into different social groups within a given context. Those categorisations might be motivated by different aspects (e.g., self-esteem and uncertainty) and can lead to inter-group bias when individuals favour members of their own group (in-group) to the disadvantage of the “out-group”. Finally, those inter-group biases can manifest in the form of prejudice, discrimination, and stereo-
Section 2.4.3 highlights more details about the social context and other subjective aspects of bias.

### 2.4.3 Subjective aspects of bias

While SIT provides a framework that suggests how inter-group bias emerges from social categorisations, the boundaries of a social group are often rather obscure and fluid. Indeed, Turner et al. (1994) argued that:

>categorising is inherently comparative and hence is intrinsically vari-
able, fluid and relative to a frame of reference. It is always context-dependent. Self-categories do not represent fixed, absolute properties of the perceiver, but relative, varying, context-dependent properties (p. 7).

Accordingly, the extent to which individuals perceive themselves as part of a given social group depends on their particular frame of reference. Thus, determining what is biased against (or in favour of) a particular group depends on the subjective nature of the group boundaries, i.e., how the perceiver categorises such a group.

Furthermore, in addition to the fluid perception of a social group, determining what is actually biased might comprise a judgement of what is right or wrong in society, i.e., a moral judgement (Waldmann et al., 2012). Therefore, the extent to which something is considered biased can be linked to cultural aspects (Brewer & Yuki, 2007; Chiu et al., 1997; Dovidio & Gaertner, 2010). For instance, Brewer and Yuki (2007) showed that culture is a key factor that contributes to shaping the meaning of in-groups, and the nature of inter-group biases.

In line with that moral aspect of bias, a notable highlight is the fact that inter-group bias can also be positive. In the context of education, for example, researchers found evidence that positive bias could be beneficial in certain circumstances (Bader et al., 2019; P. Ferguson, 2011; Harber, 1998; Harber et al., 2012). To illustrate, Bader et al. (2019) analysed 126 feedback notes of teachers and found that students were more positive towards teacher praise, which showed positive feedback as an effective way of assessment.

Those subjective aspects of inter-group bias stress the complexity of identifying such biases and, therefore, motivated this PhD project to focus on student perspectives of what they think might be negatively biased. Details about bias in educational settings and its impact on students are presented in the next section (Section 2.5). Section 2.5.1 highlights potential sources of bias, and Section 2.5.2 presents implications of bias for learning.
2.5 Bias in online learning

Assumption 3: Inter-group bias is also present in online learning and, when not addressed, it can negatively affect students in different ways, for example, by diminishing their academic performance, and by misleading key decisions when such decisions are based on biased learning materials.

As indicated above (Section 2.4), inter-group bias is directly related to group dynamics, in particular to group categorisation. Accordingly, because of social interactions, those biases can be present in a myriad of social settings, e.g., in court judgements (Gazal-Ayal & Sulitzeanu-Kenan, 2010), recruitment processes (Koval & Rosette, 2021), social media (Bliuc et al., 2018), healthcare (FitzGerald & Hurst, 2017), and educational settings (Mahmud, 2020; Quintana & Mahgoub, 2016; Vanden Bergh et al., 2010; Warikoo et al., 2016). As the context of the present research project is online learning settings, this section will cover inter-group bias in online learning platforms. The next section presents potential sources of inter-group bias in online learning platforms. Then, key implications of such biases for learning and students are presented.

2.5.1 Sources of bias

Researchers have found different biases in several aspects of learning platforms. For example, it was already shown that biases can be present in interactions between students (Bettinger et al., 2016), interactions between teachers and students (R. Baker et al., 2018), in the technology itself (Mehrabi et al., 2021), and in learning materials like textbooks (Skopec et al., 2021). Those biases can be introduced both during the early stages of learning platform development and when the course is already online and undertaken by students. Figure 2.3 summarises potential sources of bias during the life cycle of online learning platforms. More details are also presented below.
Potential biases introduced during the platform design

Potential biases can be introduced during the early stages of technology development. Two notable examples are (i) biases introduced by stakeholders responsible for authoring new learning resources like course designers, content creators, and book authors; and (ii) biases introduced in the technology itself, for example, biases in algorithms. Below, more details about each of these two aspects are provided.

As soon as new learning materials are under development, potential biases can be introduced. To illustrate, several researchers have indicated their concern about different biases across textbooks and course curricula (e.g., Skopec et al. (2021)), which was highlighted in a recent literature review (Shahjahan et al., 2022) that analysed 207 articles related to curriculum decolonisation in Higher Education (HE). The authors argued that “curriculum decolonisation” refers to

recognising the constraints placed by mono-cultural perspectives or hierarchies in one’s discipline, institution, profession, policies, and/or broader society (p. 83).

As this concept is based on such “mono-cultural perspectives”, which may be
roughly generalised as the perspectives of White Europeans, those perspectives represent in-group-out-group bias. In that case, the bias refers to the fact that the curriculum is being framed in a way to represent the views of White individuals to the detriment of the perspectives of stigmatised social groups (e.g., Black females). Therefore, when those curricula are implemented in online learning platforms, those biases (or mono-cultural perspectives) also become present in online learning.

Bias can also be incorporated into technology. Although biases in computer systems had already been investigated several years ago (e.g., Friedman and Nissenbaum (1996)), the topic has gained special attention in recent years with evidence pointing to the presence of biases in complex artificial intelligence models (Leavy et al., 2020; Mehrabi et al., 2021; Ntoutsi et al., 2020; Panch et al., 2019). The term “algorithmic bias” has been commonly used to refer to such biases, which Kordzadeh and Ghasemaghaei (2022) defined as a phenomenon that happens when:

"the outputs of an algorithm benefit or disadvantage certain individuals or groups more than others without a justified reason for such unequal impacts" (p. 1).

As with other social issues (e.g., discrimination, stereotypes), algorithmic bias can also be found in educational technologies (R. S. Baker & Hawn, 2021; Kizilcec & Lee, 2022; H. Smith, 2020). To illustrate, in a recent review of the literature (R. S. Baker & Hawn, 2021), the authors pointed out that previous studies found algorithmic bias in different aspects of learning platforms, for example, in predictive models trying to forecast student dropouts and failing rates, in automated essay scoring, and assessments of language proficiency. In contrast to those biases introduced during the platform development, other biases can be introduced in online learning platforms during the time students are already using them and consuming their content. More details about the latter are discussed in the next section.
CHAPTER 2. BACKGROUND 2.5. BIAS IN ONLINE LEARNING

Potential biases introduced when the platform is live/online

Bias in online learning platforms can also be introduced when the platform is already being used by students and instructors. It can result from interactions between the users of the platform (e.g., instructor-student, student-student) or even as a result of analytics based on biased data and algorithms. Below, those sources of bias are highlighted.

To illustrate, a recent study (R. Baker et al., 2022) tested the presence of bias in comments extracted from discussion forums of more than one hundred massive open online courses (MOOCs) and found that, on average, instructors consistently privileged White males, i.e., students from such group were 94% more likely to receive a response from their instructors than students from other groups. Other researchers (Bettinger et al., 2016) also found evidence suggesting that students from an online course were more likely to interact with peers from a similar background. Those findings reinforce biases resulting from human interactions already found in traditional educational settings (e.g., İnan-Kaya and Rubie-Davies (2022)) for which, so far, there is limited evidence in online learning.

In addition to biases present in human-to-human interactions, researchers have also identified biases resulting from decisions informed by analytical mechanisms, such as AI-based technologies (Holmes et al., 2021; Slade & Prinsloo, 2013). As indicated previously (Section 2.5.1), bias can be introduced at the time computational models are developed, which can ultimately affect the decisions of the platform’s users (e.g., teachers).

In summary, this and the previous sections outlined how bias can be introduced in online learning platforms at different stages of their life cycles. For example, during the platform development stage and when the platform has already been deployed. When those biases are not removed or mitigated, and students are exposed to them, the literature shows that they might affect the learning process in different ways. In
the next section, details about the implications of such biases for students and for their learning are presented.

2.5.2 Implications for learning

While in certain circumstances inter-group bias can have a positive impact on students, such as with positive feedback (see Section 2.4.3), several researchers have shown that inter-group bias can harm students of different social groups in a variety of ways. For example, there is evidence showing the negative impact of bias on females (Anthony, 2004; Banks, 1988; Berg et al., 2018; Spencer et al., 1999), individuals from ethnic minority groups (Fries-Britt & Turner, 2001; Jacoby-Senghor et al., 2016; Kellow & Jones, 2008; Stauffer & Buckley, 2005), disabled students (Koretz & Barton, 2004; Rosenbaum & Massey, 2005; Rousso, 2003), and students with vulnerable socioeconomic status (Bernard & Clarizio, 1981; Haller & Davis, 1980). Alongside those investigations, researchers have reported issues linked to anxiety, depression, apprehension, and motivation. Next, some of those implications are covered.

Researchers have reported several consequences of inter-group bias, e.g., anxiety, depression, negative thinking, lack of motivation, and low academic performance. To illustrate, Albuquerque et al. (2017) asked participants to take a mathematics test in a biased game-based learning environment, i.e., they manipulated the platform graphical interface to show a particular gender as dominant (e.g., by showing participants a leaderboard with only males, by providing only male avatars as choices). The authors reported that, by the end of the study, there were statistically significant changes in the levels of anxiety for female students who took the test. Similarly, a meta-analysis conducted by Lorant et al. (2003) identified a statistically significant association between socioeconomic status and depression. The analysis covered more than 50 studies, and the authors urged the scientific community to develop strategies to tackle social inequities.
To investigate the impact of negative thinking, Jordano and Touron (2017) conducted an experiment where they asked adult participants to take a working memory task. The activity comprised reading biased news about age-related memory changes (i.e., the researchers manipulated the news in a way that for one group of participants, memory changes were shown as a consequence of lifestyle, and for the other group memory changes were shown as biologically-driven). The results showed that such manipulations in the environmental context triggered participants’ concerns, and bolstered mind wandering, which had been reported earlier in the literature as a potential cause of negative thinking (Poerio et al., 2013).

Regarding student motivation, Doubé and Lang (2012) investigated students in a programming subject with an approximately equal number of female and male participants. They found that, while students had a similar extrinsic motivation to achieve their goals, the motivation of female participants was impacted negatively when exposed to the stereotype regarding the participation of girls in programming, i.e., that programming is not suitable for girls (Master et al., 2017). The authors also showed that group bias can lessen student performance.

Pennington et al. (2016) reviewed several studies that showed bias affecting females and African-Americans triggered psychological mechanisms like anxiety and negative thinking, which affected student academic performance. H. Nguyen and Ryan (2008) conducted a meta-analysis regarding the impact of gender and ethnic biases on student performance and found that explicit bias (e.g., clearly favouring particular students in a test) affected people from ethnic minority groups more intensely. The findings confirmed earlier studies dated more than two decades ago, e.g., MeCornack and McLeod (1988), Steele (1997), and Spencer et al. (1999).

This evidence shows that inter-group biases can harm students in different ways. However, before dealing with those issues, it is necessary to first identify to what extent a given environment (e.g., a course) contains bias. These reflections led to the next part of this literature review which focused on potential approaches aiming
CHAPTER 2. BACKGROUND  2.6. BIAS IDENTIFICATION

to identify bias. The next section summarises those approaches.

2.6 Strategies aiming to identify bias

Assumption 4: Despite existing efforts to understand and avoid biases in non-educational contexts, mechanisms to address ethnic biases in online learning materials and, in particular, mechanisms to address bias within large datasets are limited.

2.6.1 Non-automated strategies

Several studies have used manual - or non-automated - strategies to identify bias in learning (and non-learning) settings. Below, some of them are highlighted, which are organised into two groups: experimental approaches, which comprise scientific experiments conducted in laboratories; and linguistic-based approaches which comprise corpora analysis in which discourse analysis techniques were used to identify bias in text.

One notable example is the identification of bias using the Implicit Association Test (IAT), a laboratory-based approach that uses a scale to identify the extent to which an individual is biased against a particular group (e.g., females). The scale is part of Project Implicit, “a non-profit organisation and international collaboration of researchers who are interested in implicit social cognition” (Team, 2011). According to Project Implicit, the IAT focuses on measuring “the strength of associations between concepts (e.g., black people, gay people) and evaluations (e.g., good, bad) or stereotypes (e.g., athletic, clumsy)”. Since its development, several variations have been designed, which is evidenced in a meta-analysis that reported more than 100 studies using the IAT (Greenwald et al., 2009). Despite its popularity, the IAT has only been used to estimate the extent to which individuals are biased against particular groups rather than assessing bias in the language (e.g., it cannot spot biases
already introduced by humans in text, media, etc.).

Aiming to identify potential biases in language, researchers have proposed approaches informed by methods from Linguistics. A notable example is the use of corpus linguistics methods to identify biases in text (for example, P. Baker (2012)). To illustrate, Maass (1999) proposed a linguistic model to identify bias in language based on language abstraction. For instance, they indicate that “a positive behaviour displayed by an in-group member will be described in relatively abstract terms, whereas the same behaviour shown by an out-group member will be described in relatively concrete terms”. Therefore, the model estimates bias based on the number of abstract and concrete terms present in each text. Despite the promising results of such an approach in identifying potential biases in particular corpora, it is challenging to apply it to a large corpus (e.g., online texts) as it would require a substantial number of people to analyse the text.

Analytical techniques recently developed for the so-called “Big Data” (du Boulay et al., 2018) raise the possibility of automating the identification of biases in text. The next section provides more details about existing approaches that have been used in the attempt to identify bias in text.

2.6.2 Automated strategies

Each section of this theoretical background holds its own importance for this thesis. However, this specific section, which focuses on identifying automated approaches for detecting inter-group bias in text, serves as the foundation for Study 1 and thus underwent a more rigorous systematic process. Accordingly, it followed the rationale proposed by Wohlin et al. (2012) for conducting systematic literature reviews. This aimed at avoiding study-selection bias and comprised of defining a protocol, a checklist for assessing study quality, and data collection criteria for selected studies (see Appendix A.2 for full details).

After this rigorous and systematic process, out of 84 initial studies found across six
databases (also detailed in Appendix A.2), only two met the selection criteria:

- “Evidence That Gendered Wording in Job Advertisements Exists and Sustains Gender Inequality” (Gaucher et al., 2011) and

- “Automatically Neutralizing Subjective Bias in Text” (Pryzant et al., 2020)

The remaining 82 studies were excluded for various reasons, such as not proposing a bias identification approach, lacking peer review, or not offering an automated bias identification method. Below, they are presented in more detail.

Gaucher et al. (2011)’s strategy focused specifically on gender bias, and their approach involved a dataset of feminine and masculine words informed by previous literature (Bartz & Lydon, 2004; Hoffman & Hurst, 1990; Rudman & Kilianski, 2000) (an excerpt of these words is provided in Table 2.1). In particular, Gaucher et al. (2011) looked at potential gender bias in job advertisements of gender-dominated occupations like plumber and hairdresser. According to this approach, the presence of those words in a text corpus does not necessarily imply gender bias if the number of gender-related words is equal for females and males. In contrast, if most gender-related words are predominantly masculine or feminine, that piece of text would be gender-biased according to the authors. To illustrate, the sentence “You are confident and ambitious in these competencies” would be considered biased by their approach because the number of masculine and feminine words is different (in this case, there are three masculine words: “confident”, “ambitious”, and “competencies”; and absence of feminine words).

Table 2.1: Excerpt of gender-specific words coded by Gaucher et al. (2011)

<table>
<thead>
<tr>
<th>Feminine words</th>
<th>Affectionate, child*, cheer*, commit*, communal, compassion, connect, considerate, cooperat*, depend, emotiona*, empath*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Masculine words</td>
<td>Active, adventurous aggress*, ambitio*, analy*, assert, athlet*, autonom*, boast, challeng*, compet*, confident, courag*</td>
</tr>
</tbody>
</table>

* words that start with this form, e.g., “child*” may refer to “child”, “children”, and “child-care”.

In addition, Gaucher et al. (2011) used a questionnaire to validate their approach.
They asked English-fluent students from an introductory psychology class to analyse job advertisements according to a questionnaire about job appeal and belongingness. The advertisements were manipulated to include gender-coded words (i.e., words that were more frequent in job listings for gender-dominated occupations). Afterwards, the students were asked to rate the advertisements in terms of their sense of identity with the position being listed. As the researchers kept track of both participant gender and manipulated advertisements, they used such data to investigate the relationship between those two variables.

In contrast, Pryzant et al. (2020)’s approach focused on identifying and neutralising potential subjective biases related to demographic variables like ethnicity and gender. Subjective bias in text comprised subjective language that makes a piece of text skewed based on opinion, feeling, or taste (Pryzant et al., 2020). Some examples of subjective bias identified by Pryzant et al. (2020) are shown in Table 2.2.

<table>
<thead>
<tr>
<th>Biased sentence</th>
<th>Neutral sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>The authors’ exposé on nutrition studies.</td>
<td>The authors’ statements on nutrition studies.</td>
</tr>
<tr>
<td>He started writing books revealing a vast world conspiracy.</td>
<td>He started writing books alleging a vast world conspiracy.</td>
</tr>
<tr>
<td>Jewish forces overcome Arab militants.</td>
<td>Jewish forces overcome Arab forces.</td>
</tr>
<tr>
<td>The lyrics are about mankind’s perceived idea of hell.</td>
<td>The lyrics are about humanity’s perceived idea of hell.</td>
</tr>
<tr>
<td>Marriage is a holy union of individuals.</td>
<td>Marriage is a personal union of individuals.</td>
</tr>
</tbody>
</table>

To illustrate, the sentence “Jewish forces overcome Arab militants” is considered biased by Pryzant and colleagues because the word “militants” is adding a negative association to Arab forces, i.e., the sentence is skewed in favour of Jewish forces rather than being presented in a neutral way (e.g., “Jewish forces overcome Arab forces”). Another example is highlighted in the sentence “The lyrics are about mankind’s perceived idea of hell”, where the word “mankind” alludes to men rather than being neutral in terms of gender (e.g., “The lyrics are about humanity’s perceived idea of hell”).

To identify such biases, Pryzant et al. (2020) used natural language processing (NLP) techniques to train two models to predict bias in sentences and suggest an
equivalent neutral sentence. Then, a mixed strategy of quantitative and qualitative methods was used to validate such models. For the quantitative evaluation, a set of metrics were used to evaluate their algorithm in which the researchers reported a 93.52% performance and 46.8% accuracy. For the qualitative evaluation, they hired English-speaking crowd workers and asked them to label a sample of sentences as biased or neutral. Then, the responses from the participants were compared to the algorithm’s predictions. According to Pryzant and colleagues, their ML approach performed better than less experienced Wikipedia editors.

Regarding the research context, both studies were conducted in online settings. Gaucher et al. (2011) conducted their study in the context of job advertisements, where their dataset was extracted from a job posting website at the University of Waterloo, Canada. Pryzant et al. (2020) performed their study by using substantial data from Wikipedia pages. In addition, they also used datasets from books and magazine articles, headlines of partisan news articles that they identified as biased, and sentences from speeches of a US politician. More details about the methodological aspects of each of those approaches are provided in Section 4.2.2.

In the next section, the key limitations of those approaches are highlighted.

### 2.6.3 Limitations of existing strategies

Two main limitations were identified in the reviewed approaches: they did not consider complex instances of biases, and they did not account for the identification of bias in large educational datasets.

Despite the methodological robustness of both approaches, they were both narrow in the sense that they aimed to identify bias at a word level, which may perform well for a basic instance of bias (e.g., cursing words) but might not be effective for complex ones (e.g., when the context of a sentence needs to be considered). For instance, those approaches do not take the broader context into consideration (e.g., sentence, paragraph, document, course).
While several studies have focused on the algorithmic bias in general (e.g., Kordzadeh and Ghasemaghaei (2021) recently reviewed over 100 articles), it seems that the automatic identification of inter-group bias in educational texts may not have received much attention from the research community. As noted above, only two approaches were found during this stage of the literature review. That limited evidence regarding the performance of those approaches, when applied to online learning settings, led to the first research question of this PhD thesis:

**RQ1.0: How do computational approaches, that claim to uncover inter-group bias in non-educational texts, perform in text-based online learning materials?**

In essence, that research question was drawn based on the hypothesis that better understanding the limitations of existing approaches was an initial step towards the development of a more effective way to uncover potential biases within online learning materials. Indeed, understanding the research problem is considered by many researchers as a primary step in the research process (Babbie, 2020; Creswell & Creswell, 2017; Kumar, 2018). Accordingly, that research question was covered during the initial stages of this PhD project and was the guiding question for Study 1 (Chapter 4).

Furthermore, another challenge identified during this literature review and upon the results of Study 1 was the limited evidence of how ethnic bias manifests in learning materials. For example, Section 2.4.3 showed that bias can be subjective (i.e., what is considered biased might depend on specific perspectives and contexts). In addition, as shown above, the two identified approaches were not designed for educational settings, which means they did not provide evidence of how inter-group bias manifest in learning texts. Those limitations led to the second research question (which guided Study 2 – Chapter 5):

**RQ2.0: How does bias against certain ethnic groups manifest in text-based online learning materials?**
Despite those limitations, automating the identification of bias in texts seems promising for handling the enormous amount of online data. In online learning platforms, one field that has gained special attention from the research community for dealing with complex educational issues (and large amounts of data) is Learning Analytics. The next section highlights how Learning Analytics have been used to address social issues in online learning and why it makes sense for this study.

2.7 Learning analytics

Assumption 5: Learning Analytics seem promising for addressing subjective and complex issues in online learning.

As indicated above, social issues like biases have been incorporated into online learning platforms, which can impair students in different ways. However, one key issue for identifying biases in online learning platforms is the massive amount of data present in such settings, which along with the subjectivity of bias makes the task unpractical for being executed by hand. With the aim of improving learning processes, a notable field has drawn attention from the research community over recent decades: Learning Analytics (LA), which can be defined as:

_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).

Spears (2007) suggested that the growth and expansion of LA were caused by the increasing number of educational data available online that resulted from the emergence of online learning platforms. As a research area, LA combines a variety of fields (e.g., educational data mining, web analytics, business intelligence), which brings together the perspectives of different researchers and practitioners (R. Ferguson, 2012).
In terms of objectives, Rienties et al. (2020) argued that LA aims at improving learning processes and “focuses mostly on instruction and communication, student learning objectives and natural language processing.” The next section highlights some studies that used LA to address complex issues in online learning platforms.

2.7.1 LA and complex issues in online learning

The use of LA to address or mitigate complex issues in online learning platforms is not new. For instance, researchers have used LA to predict performance (Namoun & Alshanqiti, 2020), understand student emotions and behaviour (Sedrakyan et al., 2020), and uncover performance gaps (Q. Nguyen et al., 2020). Below, recent findings providing evidence for those aspects are presented.

Perhaps, the most notable attention LA has gained is its use for making predictions regarding student success or failure in online courses. To illustrate, a systematic literature review that covered 10 years of research (2010-2020) regarding predicting student performance (Namoun & Alshanqiti, 2020) found 62 articles related to the topic. They indicated that most studies used regression and supervised machine learning models to classify student performance. The review also showed that several models were trained based on student academic emotions, i.e., parameters based on student motivation and interest.

In addition to those findings, there is also evidence highlighting positive aspects of LA towards the improvement of learning design (Mangaroska & Giannakos, 2018; Rienties et al., 2017), and to support administrative decisions based on robust measures of efficiency and effectiveness of organisational processes in educational settings (Balica et al., 2018; Jones, 2012). Another study (Rienties & Toetenel, 2016a) used multiple regression models to investigate the impact of learning design on 111,256 student satisfaction and performance. Their findings stressed the importance of taking learning design into account to better understand student performance in both blended and online settings.
Other uses of LA have focused on predicting student behaviours (Dooley & Makasis, 2020; Sedrakyan et al., 2020). For instance, (Dooley & Makasis, 2020) applied LA to understand student learning behaviours in flipped classrooms and found that early access to learning materials was related to better performance. Then, they used such findings to improve the course design. In another example (Sedrakyan et al., 2020), LA was used to understand the relationship between feedback to learners (through LA dashboards) and the regulation of learning. The results suggested that dashboards informed by LA had a positive impact on the way students regulated their learning (Sedrakyan et al., 2020).

LA has also been used to understand performance gaps in online learning platforms (Q. Nguyen et al., 2020; Sabnis et al., 2022). For example, Q. Nguyen et al. (2020) found evidence indicating that individuals from an ethnic minority background (e.g., Black and African students) were at more risk of not completing their online course successfully. A more recent study (Sabnis et al., 2022), also about LA to uncover discrepancies in student performance, looked for differences in procrastination behaviour across socio-demographic groups and found “higher levels of procrastination behaviour among males, racial minorities, and first-generation college students than their peers” (Sabnis et al., 2022, p. 133). Although those studies did not address potential biases in learning materials, they highlight the potential of LA in addressing inequities in online learning platforms.

Those promising results regarding the use of LA to solve complex issues in online learning platforms were a key motivator for adopting LA in this PhD project. In particular, the fact that LA takes into account large datasets from online learning settings means that such techniques might be able to assist in the identification of inter-group bias in such settings. That motivation led to the third research question of this thesis:

**RQ3.0: To what extent can Learning Analytics assist in the identification of**
ethnic bias in text-based online learning materials?

Accordingly, that research question guide Study 3 of this PhD project, which is presented in Chapter 6. Furthermore, while particular details about each study are presented in their respective chapters (i.e., in chapters 4–6), the methodological aspects of this thesis are covered in the next chapter (i.e., Chapter 3: Methodology).

2.8 Chapter summary

In this chapter, the key theories on which this PhD project is grounded were presented. For instance, fundamentals of online learning and broader ongoing challenges like inequities were initially explored, which led to the assumption that those challenges make online learning less inclusive and unfair for certain students. Afterwards, the fundamentals of inter-group bias were discussed, which covered the Social Identity Theory and subjective aspects of bias. Then, once key aspects of online learning and inter-group bias were addressed, the subsequent section focused on the sources of bias in online platforms and the impact of such biases on students. Subsequently, existing strategies aiming at the identification of bias were reviewed, and their limitations were highlighted. Finally, this chapter covered key fundamentals of LA, in particular, how it has been used to address complex issues in online learning settings, which makes it a promising candidate for dealing with ethnic bias in online learning texts.

While this chapter covers the theoretical basis for the present research project, the next chapter focuses on its methodological basis and underlying research methods. Drawing on this theoretical basis, in this study LA is used to collect and analyse data from Higher Education students and text-based online learning materials. In addition, the present PhD project also uses natural language processing techniques to identify potential biases in such materials.
3 | Methodology

3.1 Chapter overview

Chapter 2 highlighted the theoretical basis for this research, which includes how each research question emerged from the literature and the research aims. The initial sections of the present chapter cover the research philosophy (Section 3.2) and the underlying methodology (Section 3.3) that supported the operationalisation of those research questions. Particular details of the procedures and the materials used in each research study will be covered in the subsequent chapters (Chapters 4-6). Once the research paradigm and underlying methodology are covered, Section 3.4 presents the research design, i.e., the studies and research cycles entailed by this doctoral project. In the end, this chapter highlights ethical and data protection issues related to the execution of this methodology (Section 3.5).

As discussed in the previous chapter, the core research questions covered in this thesis are:

- RQ1.0: How do computational approaches, that claim to uncover inter-group bias in non-educational texts, perform in text-based online learning materials?
- RQ2.0: How does bias against certain ethnic groups manifest in text-based online learning materials?
- RQ3.0: To what extent can Learning Analytics assist in the identification of ethnic bias in text-based online learning materials?

3.2 Research philosophy

Saunders et al. (2019) summarised the definition of research philosophy as “a system of beliefs and assumptions about the development of knowledge”. Those assumptions are made during different stages of a research project (Burrell & Mor-
gan, 2017), and they shape how a researcher interprets the wider literature, respective context, research questions, methods, and findings (Crotty, 1998). In addition, Saunders et al. (2019) indicated that a well-defined set of assumptions contributes to a consistent research philosophy, which grounds the methodological choice, research strategy, and the methods behind the data collection and analysis. Therefore, the research philosophy behind this doctoral project is highlighted before presenting the methodological approach in this chapter.

Although there are different ways to frame and classify a research philosophy, several researchers have considered three aspects to classify the assumptions underlying a research philosophy: ontology, epistemology, and axiology. Those aspects are briefly summarised below.

### 3.2.1 Ontology, Epistemology, and Axiology

Ontology comprises the nature of beliefs about reality (Richards, 2003), and it helps understand the facets that constitute the known world (Scott & Usher, 2011). Researchers may have implicit and explicit assumptions about reality (Rehman & Alharthi, 2016), where Patton (2002) suggested that studies are usually framed according to two different perspectives of reality: (i) “a singular, verifiable reality and truth”; and (ii) multiple realities that are socially constructed based on different views and contexts.

Epistemology is related to the nature of knowledge and how it is acquired and validated Gall et al. (1996). Accordingly, four sources of knowledge have been proposed (Kivunja & Kuyini, 2017): (i) intuitive knowledge, where the assumptions rely on beliefs, faith, and intuition; (ii) authoritative knowledge, when the assumptions are based, for example, on books, organisations, and leaders; (iii) logical or rationalist knowledge, when the truth is determined by a reasoning and logical process; and (iv) empirical knowledge, when the emphasis relies on experience, observation, or experimentation.
Axiology comprises the role of values and ethics behind a research project (Saunders et al., 2019), e.g., to what extent a research project will reflect the researcher’s or participants’ values. Kivunja and Kuyini (2017) - based on the work conducted by Mill (2016) - suggested four criteria to guide those values and principles namely, teleology, deontology, morality, and fairness. Particularly, teleology “refers to attempts made in research to make sure that the research results in a meaningful outcome that will satisfy as many people as possible” (Kivunja & Kuyini, 2017), which is related to what is morally good and right when conducting a research project. Deontology concerns the consequences of actions taken during a research project, and how they might affect stakeholders, e.g., other researchers, participants, and the public (Scheffler, 1994). Morality “refers to the intrinsic moral values that will be upheld during the research” (Kivunja & Kuyini, 2017), e.g., the researcher’s responsibility and truthfulness when recruiting participants. Finally, Kivunja refers to fairness as “the researcher’s attention to the need to be fair to all research participants and to ensure that their rights are upheld”.

In summary, establishing a research philosophy depends on a diverse set of assumptions involving personal reflection and exploration (Alvesson & Sköldberg, 2017; Haynes, 2012). As such aspects can lead to a multitude of perspectives, a single research project can be positioned within a range of research philosophies. Despite that variety of possibilities, researchers have considered basic sets of assumptions to frame their research projects. Those taken-for-granted assumptions are usually referred to as research paradigms. Kivunja and Kuyini (2017) summarised the definition of paradigm as positions about epistemology, ontology, and axiology, which can influence the methodology of a research project. Table 3.1 provides a summary of beliefs, assumptions, norms, and values of five common paradigms.
### Chapter 3. Methodology

#### 3.2. Research Philosophy

<table>
<thead>
<tr>
<th>Ontology</th>
<th>Epistemology</th>
<th>Axiology</th>
<th>Typical methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positivism</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Critical realism | Epistemological relativism. Knowledge historically situated and transient. Facts are social constructions. Historical causal explanation as contribution. | Value-laden research. Researcher acknowledges bias by world views, cultural experience and upbringing. Researcher tries to minimise bias and errors Researcher is as objective as possible. | Retrospective, in-depth historically situated analysis of pre-existing structures and emerging agency. Range of methods and data types to fit subject matter. |
| Stratified/layered (the empirical, the actual and the real). External, independent. Intransient. Objective structures. Causal mechanisms. | | | |

| Interpretivism | Theories and concepts too simplistic. Focus on narratives, stories, perceptions and interpretations. New understandings and worldviews as contribution. | Value-bound research. Researchers are part of what is researched, subjective. Researcher interpretations key to contribution Researcher reflexive. | Typically inductive. Small samples, in-depth investigations, qualitative methods of analysis, but a range of data can be interpreted. |
| Complex, rich. Socially constructed through culture and language. Multiple meanings, interpretations, realities. Flux of processes, experiences, practices. | | | |

| Postmodernism | What counts as 'truth' and 'knowledge' is decided by dominant ideologies. Focus on absences, silences and oppressed/repressed meanings, interpretations and voices. Exposure of power relations and challenge of dominant views as contribution. | Value-constituted research. Researcher and research embedded in power relations. Some research narratives are repressed and silenced at the expense of others. Researcher radically reflexive. | Typically deconstructive - reading texts and realities against themselves. In-depth investigations of anomalies, silences and absences. Range of data types, typically qualitative methods of analysis. |
| Nominal | Complex, rich. Socially constructed through power relations. Some meanings, interpretations, realities are dominated and silenced by others. Flux of processes, experiences, practices. | | |

| Complex, rich, external. ‘Reality’ is the practical consequences of ideas. Flux of processes, experiences and practices | | | |

#### 3.2.2 The present project within a pragmatic paradigm

To establish the research philosophy behind this project and, therefore, make a more informed choice of a research paradigm, two aspects suggested by Saunders et al. (2019) were considered: (i) the researcher’s beliefs and assumptions regarding the research topic; and (ii) the research philosophies commonly used within the research area. The next section covers how these two aspects were instantiated for the
present research by highlighting the core assumptions that informed the choice of a pragmatic paradigm.

**Beliefs and assumptions regarding the research topic**

In terms of **ontology**, the assumptions of this project are that the world is part of a universal and true reality. For example, the existence of ethnic disparities in educational settings, like the ones identified in Section 2.3.2, the issues faced by students because of inter-group bias (see Section 2.5.2), and the possibility of automating certain tasks using computers like predicting student performance and behaviours (Namoun & Alshanqiti, 2020; Sedrakyan et al., 2020) as indicated in Section 2.7.1.

Nonetheless, a notable exception is what comprises inter-group bias, as such characterisation might depend on the perceptions of particular groups of individuals. Indeed, as shown in Section 2.4.3, the characterisation of what comprises inter-group bias can be subject to contextual elements (e.g., culture) and how individuals place themselves within a given social group (Tajfel, 1970; Tajfel et al., 1971). In other words, the reality of inter-group can be perceived as relative and subject to different interpretations (see also Turner et al. (1994)). Accordingly, as will be shown in Chapter 5, the second study of this research project looked at how students perceive ethnic bias in text-based learning materials.

Regarding **epistemology**, this project focused on answering a set of research questions with the goal of contributing to the mitigation of ethnic issues in educational settings. That aspect makes the present project problem-focused and, therefore, it concentrated on the practical meaning of knowledge in specific contexts. For example, the metrics used to validate several machine learning models\(^1\) were considered an acceptable source of knowledge because they were widely used in the context of computational modelling (see also Section 2.7 for more details). In contrast,

\(^1\)See Chapter 6 for more details about the machine learning models used in this PhD project.
the perspective of students regarding what they perceive as inter-group bias was also considered a valid source of knowledge given that what is considered biased against a social group can depend on how individuals place themselves in that group (see Section 2.4).

**Axiology.** Although the research topic was chosen based on particular researcher values (e.g., equity education and social justice), this project involved different reflective moments aiming at the mitigation of any potential negative impact on students. A notable exception was during the data analysis, where the researcher tried to detach from the object of study in order to prevent potential researcher biases. The APA Dictionary of Psychology defines researcher bias as “*any unintended errors in the research process or the interpretation of its results that are attributable to an investigator’s expectancies or preconceived beliefs.*” (Association, 2023). Nonetheless, most parts of the present project were value-driven.

Although those beliefs and assumptions regarding the research project could already be framed within a pragmatic paradigm, further evidence was sought by looking at commonly used paradigms related to the research areas.

**Research paradigms within the research area**

As highlighted in Chapter 2 (Theoretical Background), inter-group bias is a subject widely studied in Social Psychology (e.g., Greenwald and Krieger (2006), Pennington et al. (2016), and Tajfel (1970)), while online learning is usually investigated by researchers from Computer Science or Education (e.g., Allen and Seaman (2007), Keller (2005), and Rienties and Jones (2020)). That makes the essence of this thesis extend across those different areas as its general objective involves inter-group bias in online learning.

Accordingly, as each of those disciplines has different approaches to conducting research, different research paradigms can be found across those areas. For example, some researchers from Social Psychology might choose Interpretivism to
investigate the subjective experiences of individuals in their social contexts (J. A. Smith, 2015). In contrast, researchers from Education and Computer Science, who might be more interested in emphasising practical problem-solving and the use of multiple methods to investigate complex issues, might opt for a pragmatic paradigm (Creswell & Creswell, 2017; Shneiderman, 2016). Therefore, the intersection of research philosophies from multiple disciplines was another factor informing the choice of a pragmatic paradigm for this doctoral project.

3.2.3 Methodological choice

By definition, a pragmatic paradigm uses the research methods that are more appropriate for answering the research questions (Kivunja & Kuyini, 2017), which gives the possibility of choosing from different methodologies. For example, methodologies predominant in a positivist paradigm (e.g., experimental, correlational, randomised control trials, survey research), but also methodologies widely used within other paradigms like a case study, phenomenology, ethnography, and action research. Accordingly, this PhD project is framed within a design-based research (DBR) approach (see Section 3.3.1), which gives the possibility of combining multiple of those methodologies in the scope of a single research project. The next section provides more details about the DBR approach and how it fits the scope of this project.

3.3 Underlying methodology: Design-Based Research (DBR)

In this section, the underlying methodology that guided the development of this project is presented, i.e., design-based research (DBR). First, its fundamentals are presented, which are followed by its benefits and some examples of applications in educational technologies. Furthermore, this section also highlights criticisms DBR
has received and indicates how DBR fits the scope of this PhD project.

### 3.3.1 Fundamentals of DBR

DBR can be defined as “an important methodology for understanding how, when, and why educational innovations work in practice (...), [which] blends empirical educational research with the theory-driven design of learning environments” Baumgartner et al. (2003, p. 5). Although that definition presents DBR as one single methodology, some researchers focused on its flexibility and referred to DBR as a set of approaches that allows the usage of multiple methodologies. For example, Barab and Squire (2004) defined DBR as “a series of approaches, with the intent of producing new theories, artefacts, and practices that account for and potentially impact learning and teaching in naturalistic settings” (p. 2). In particular, DBR is usually described as an emerging approach derived from design research.

Nonetheless, despite emerging from educational design research, researchers have stressed key differences between DBR and other educational methods. Usually, they claim that traditional educational research aims at investigating the efficacy of educational interventions, which are often conducted in non-naturalistic settings (e.g., in controlled laboratories), while DBR is about understanding and documenting designed intervention in practice (Anderson & Shattuck, 2012; Baumgartner et al., 2003). This practical aspect of DBR makes it “a practical research methodology that could effectively bridge the chasm between research and practice in formal education” (Anderson & Shattuck, 2012, p. 16). In particular, Baumgartner et al. (2003, p. 5) proposed five core principles that distinguish DBR from other methodologies, and that should be accounted for when using DBR:

1. The design of learning environments and learning theories are intertwined. This means that interventions are based on relevant theories that aim at informing practice.

2. The research is developed through an iterative process with cycles of design,
implementation, analysis, and re-design.

3. The research should lead to theories that communicate relevant implications to educational designers and practitioners.

4. Research should consider authentic settings, i.e., interventions take place in real-world scenarios (also referred to as “naturalistic settings”).

5. Methods used along the research cycles should be reliable and connected to promote outcomes of interest.

In terms of the processes behind each cycle of DBR, McKenney and Reeves (2018, p. 83) suggested three core processes (also known as “micro-cycles”): (i) analysis and exploration; (ii) design and construction; and (iii) evaluation and reflection. As indicated above, DBR is an iterative approach, which means that a particular research project can have as many of those micro-cycles as needed. More details about each one is provided below (see also Figure 3.1 from Armstrong et al. (2022) for an overview of the iterations).

**Analysis and exploration.** The analysis comprises the problem definition, which should be guided by the research context and needs (McKenney & Reeves, 2018). In addition, the same authors suggested the analysis should focus on reviewing the literature to refine the initial perception of the problem as well as its potential causes. Furthermore, they added that the exploration may take place during the problem analysis and aims at seeking and learning “from how others have viewed and solved similar problems” (p. 85). Once the problem becomes clearer to the researcher, the design and construction begin.

**Design and construction.** This micro-cycle comprises “a deliberative-generative process that yields a well-considered intervention which is grounded in both theory and reality” (McKenney & Reeves, 2018, p. 109). They also stressed that each project design is unique, which means that “the most fruitful and fitting approaches for a specific educational design study” (p. 110) should be selected and used. Ac-
Accordingly, this stage can produce different outputs like teacher guides and software, which can be the subject of the evaluation and reflection micro-cycle (McKenney & Reeves, 2018).

**Evaluation and reflection.** Finally, this micro-cycle aims at informing the external research community in addition to driving the development of potential new interventions (McKenney & Reeves, 2018). In particular, they stressed this step should focus on empirically testing the design ideas and potential prototypes resulting from the previous micro-cycle. Then, once such testing is conducted, “the findings are reflected upon, with the aim of refining (theoretical) understanding about it, how, and why intervention features work” (McKenney & Reeves, 2018, p. 133).

![Figure 3.1: The core processes of DBR (Armstrong et al., 2022).](image)

Since its emergence, researchers have highlighted a variety of ways in which a project can benefit from that iterative process of micro-cycles. For instance, Wang and Hannafin (2005) summarised such benefits as a set of attributes that, together, makes DBR unique when compared to other approaches like design experiments (Cobb et al., 2003), or design research (Edelson, 2002). Those attributes are:

- **Pragmatic.** DBR intertwines theory and practice and makes the value of
CHAPTER 3. METHODOLOGY  3.3. UNDERLYING METHODOLOGY

theory appraised in a practical context.

• **Grounded.** The research design is driven by relevant research and theories.

• **Iterative and Flexible.** Processes are based on iterative cycles of design, implementation, and analysis. In addition, the initial research plan is revisited and adjusted after each cycle, which makes it flexible.

• **Integrative.** Different methodological perspectives are used to sustain the ongoing research based on the needs of each cycle.

• **Contextual** The research process, research findings, and research changes are documented according to the research context.

Given the advantages of DBR for educational research, various researchers have adopted such an approach in their studies (Cowling & Birt, 2018; Dolmans, 2019; Johnson et al., 2017; Koivisto et al., 2018; Sandanayake et al., 2021; Vanderhoven et al., 2016). To illustrate, some of those studies are highlighted in the next section.

### 3.3.2 DBR for educational technologies

Given its benefits, DBR has been used as a promising methodology for educational technologies. To illustrate, Johnson et al. (2017) used DBR to examine factors influencing levels of learning in an online discussion forum of undergraduate students. Given the flexible aspects of DBR like the possibility of mixing different methods in a single study, the researchers were able to combine quantitative and qualitative data analyses from three different data sources: an online survey, interviews from students and faculty, and posts from the online forum. In addition, thanks to the DBR iterative aspect, the study was developed in two research cycles that happened over two-years so that the study could cover an online course that took place at the same time of the academic years (Johnson et al., 2017).

Previous research projects have also used DBR in interdisciplinary research. For example, Koivisto et al. (2018) applied DBR in the context of Nurse Education for
developing principles of design in nursing simulation games. The authors indicated that the study comprised two research cycles, where features for a simulation game were initially defined based on the literature. Then, nursing students tested the game as part of an online learning activity and provided feedback through interviews. After analysing the interview data, the researchers concluded that

*the findings contribute not only to the development of simulation games but also to the use of design-based research methodology in developing new technology-enhanced learning environments in the field of nursing education* (Koivisto et al., 2018, p. 119).

Therefore, highlighting the application of DBR in interdisciplinary research.

In another research project (Vanderhoven et al., 2016), DBR was used for developing online learning materials about the risks of social networks. The project comprised eight iterative studies and involved more than 1500 students of secondary education. The authors also indicated that a mixed method approach was also used, which included an observational study, theoretical evaluation of existing materials, survey study, and quasi-experimental studies.

In addition to the aforementioned examples, recent research has utilised Design-Based Research (DBR) as a method for creating and enhancing learning environments (Cowling & Birt, 2018; Dolmans, 2019; Sandanayake et al., 2021; Wolcott et al., 2019). For instance, Sandanayake et al. (2021) employed DBR to develop a framework for generating open educational resources in online undergraduate courses. The authors conducted multiple studies, implementing an intervention, and using a mixed-methods approach to data analysis. In one study, feedback from 102 surveyed undergraduate students was incorporated to refine the open educational resources, which demonstrates the practical application of DBR in educational design.

Nonetheless, like any research method, DBR is not without its shortcomings, and
several challenges and critiques have been raised by other scholars. Further details are elaborated below.

### 3.3.3 Challenges and criticisms of DBR

Despite its wide range of applications and benefits, there are several challenges and criticisms related to DBR, e.g., methodological consistency, threats to validity, and feasibility for short-term projects. One major concern is the extent to which DBR projects can keep their consistency as different methodologies can be adopted. Wang and Hannafin (2005) stressed that DBR can be internally consistent but reflect different levels of discipline and rigour as it gives the possibility of combining multiple research frameworks. Nonetheless, despite such criticisms regarding the combination of multiple research methods, researchers have demonstrated that such an approach can be useful when investigating complex issues in learning technologies (Sandanayake et al., 2021; Vanderhoven et al., 2016).

For example, in the research project mentioned above about the design of learning materials (Section 3.3.2), Vanderhoven et al. (2016) used a qualitative method to gather initial insights about the existing learning materials, then conducted larger scale interventions followed by several cycles of quantitative data analyses to evaluate a revised version of such materials. In particular, the qualitative approach aimed at identifying the needs of educational stakeholders (i.e., teachers, school advisers, and learning designers) by using a focus group approach in which participants were invited to an open task. In the task, the participants were asked to write down both positive and negative aspects of existing learning materials, and their most urgent needs in terms of learning materials. Once stakeholder needs were mapped, the researchers revised the learning materials and conducted subsequent larger-scale interventions to evaluate the improved version of the materials. For instance, students were surveyed in most of those interventions, and the data analyses comprised

---

2For instance, focus group can be defined as “a research technique that collects data through group interaction on a topic determined by the researcher” (Morgan, 1996).
quantitative methods like inferential statistics.

Another challenge is the possibility of threats to validity that may be added by researchers. For instance, Barab and Squire (2004) highlighted the involvement of a researcher in the different studies of a project (i.e., conceptualisation, design, development, implementation, reflection) may compromise credible and trustworthy assertions. For example, Karsten and van Zyl (2022) used a DBR approach to investigate how a South African institution could design effective student assistance programmes by taking student contexts into account. To guarantee the validity of those studies, the lead researcher – a PhD Student – constantly interacted with the mentoring team and gathered input from a variety of students. In essence, Karsten and van Zyl (2022) involved other individuals in the research project by incorporating the feedback of stakeholders, which contributed to mitigating potential threats introduced by the lead researcher. Similarly, the present PhD project involved constant meetings with the supervisory team and adopted data from a variety of students (for details about participants, see Section 4.2.2 and Section 5.2.2).

Finally, researchers have also questioned the feasibility of DBR for short-term investigations (e.g., PhD projects) as the number of research cycles can be demanding and time-consuming (McKenney & Reeves, 2018). Despite such criticism, a variety of successful PhD projects have used DBR as an underlying research method (Abdelmawgoud, 2022; Cernusca, 2007; Holmberg, 2019; Holmes, 2013), which was also highlighted in studies that focused on investigating the suitability of DBR for PhD projects specifically (Goff & Getenet, 2017; Pool & Laubscher, 2016).

For example, Pool and Laubscher (2016) illustrated the use of DBR in a PhD project in which DBR was used to convert a face-to-face course into a blended learning course. Six research cycles were designed, which also included mixed methods like interviews and quantitative survey data. The authors highlighted that DBR should not be limited to long-term projects as it

*might reduce the scope of this methodology, and might deter students*
from applying design-based research methodology to their short-term projects such as a Masters or PhD study (Pool & Laubscher, 2016, p. 10).

Taking both benefits and limitations of DBR into account, an initial reflection at the beginning of this project concluded that the benefits would compensate for the challenges. The next section provides more details of how the present thesis fits such a methodological approach.

### 3.3.4 DBR and this research

The present research adopted a DBR rationale based on the need for a methodology that could support the complexity of inter-group bias and, at the same time, the design of a practical mechanism to support the identification of inter-group bias in learning materials. In addition, this PhD project is also in resonance with the key principles of DBR indicated above. More details are provided below.

Ford et al. (2017) suggested the use of DBR when “the specifics of the problem require assessment, clarification, and solution design” (p. 52), which is in line with this thesis’ key objectives (e.g., to understand how ethnic bias manifest in learning environments, and provide practical implications for designing a practical solution). For instance, the literature review showed that inter-group bias is linked to inter-group categorisation (see Section 2.4), where individuals may categorise themselves based on various characteristics (Tajfel, 1970; Tajfel et al., 1971). It also showed that inter-group bias, when not removed from learning environments, can negatively impact students in different ways, e.g., by impairing their academic performance (see Section 2.5.2). Therefore, those aspects suggest that inter-group bias is a complex societal problem as it requires a deeper understanding of how inter-group bias manifests in learning environments.

In addition, the key aspects of DBR described by Wang and Hannafin (2005) fit the scope of this project. First, this project was about inter-group bias, a concept
grounded on the group categorisation theory (Section 2.4). Such theoretical perspective informed the design of practical solutions for understanding and identifying inter-group bias in learning environments. Second, the overall research process needed to be integrative to support the use of different and reliable research methods, in particular, to support a pragmatic research paradigm. Indeed, Anderson and Shattuck (2012) indicated that DBR resonates with a pragmatic research paradigm given its flexibility of methods and focus on “authentic and meaningful issues”. Third, the research processes, findings, and changes were attached to the context of learning materials, in particular, text-based open educational resources. Finally, the research plan was reviewed after each research cycle to inform the next one. That cyclic process of reviewing and improving a subsequent research cycle made a flexible and iterative design essential for this PhD project. In the next section, the research design is presented, which includes more details about that iterative process.

3.4 Research design

This section covers the overall research design of this thesis. As previously mentioned, this PhD project comprised three studies, and each study was guided by one core research objective (see Section 1.3). To illustrate the overall design, Figure 3.2 depicts each study along with their respective research questions and research cycles.

While particular methodological aspects of each study are detailed in the subsequent chapters (Chapters 4–6), the essential research methods used during this PhD project can be classified into two main categories: data collection methods, and data analysis methods. Table 3.2 shows an overview of those methods along with the overall research design.

The next section (i.e., Section 3.4.1) discusses the overall rationale for those meth-
3.4.1 Key research methods

A key data collection method used during this doctoral project was web data extraction. According to Ferrara et al. (2014), that method refers to the process of using software to automate the extraction of data from web data sources (e.g., online doc-
ments, web pages, etc.). In addition, it allows the collection of data for a variety of applications that require dealing with large amounts of online data (Ferrara et al., 2014). Accordingly, Web data extraction was chosen for this PhD project because it allowed the automation of the extraction of text-based learning materials from online learning platforms. Without that automation, a huge human effort would be necessary to download each learning material by hand, which could compromise the feasibility of this project within the funded time frame. In addition, extracting data directly from the Web allowed the collection of learning materials that are used in real-world online learning platforms, which benefited the validity\(^3\) of consequent data analyses.

Furthermore, online survey was another data collection method used in this PhD project. Fink (2003) defined a survey as a method to obtain information from a sample of individuals using a set of questions. While Fink (2003) noted that surveys can be implemented in a variety of ways (e.g., face-to-face interviews, telephone questionnaires, online questionnaires), Evans and Mathur (2005) highlighted that online surveys can benefit a research project in various ways, for example, by offering convenience, cost-effectiveness, speed, flexibility, increased response rates, and the ability to reach large sample sizes. Beyond those benefits of online surveys, a primary reason that motivated the use of online surveys in this doctoral project was the possibility of reaching students from a variety of social groups, which was essential for capturing different perspectives regarding inter-group bias in learning materials. As shown in the previous chapters of this thesis (e.g., in Section 2.4.3), what is considered “biased” might vary across different groups, therefore, reaching a more diverse population of students would benefit the representativeness of those perspectives.

Regarding data analysis methods, a range of strategies was used according to the needs of each study (see chapters 4–6 for details). In essence, most of those meth-

\(^3\)According to Bolarinwa (2015), “validity expresses the degree to which a measurement measures what it purports to measure.”
ods comprised statistical data analysis, which can provide empirical evidence to support or reject hypotheses, make predictions, and inform decisions in a rigorous and objective manner (Agresti, Finlay, et al., 2009). In addition, James et al. (2013) noted that statistical methods provide a rigorous and systematic approach to analysing quantitative data, which allows researchers to identify and quantify relationships between variables, and draw conclusions based on data-driven evidence. Accordingly, statistical methods were chosen for most parts of this thesis because the data analyses conducted in each study were based on quantitative data and involved hypothesis testing and evaluation of models to predict bias.

Furthermore, in addition to statistics-based methods, Study 2 also used Thematic Analysis, which is a research method for identifying, analysing, and reporting patterns within qualitative data, enabling researchers to gain insight into the experiences and perspectives of participants (Flick, 2022). Thematic Analysis is also known for allowing researchers to generate rich, detailed, and nuanced understandings of complex phenomena, and can be applied in a wide range of research settings and disciplines Braun and Clarke (2012). Accordingly, thematic analysis was selected for this doctoral project based on its interdisciplinary nature and its need for understanding how ethnic bias manifest in text-based online learning materials while accounting for the perspectives of different ethnic groups.

The next sections present an overview of the key methodological aspects of each study. As previously mentioned, methodological details about each study are provided in the respective chapters.

3.4.2 Study 1 (see Chapter 4)

Study 1 was motivated by limited evidence regarding the suitability of existing bias-classification approaches for classifying inter-group bias in text-based learning materials (see Section 1.2). In addition, Study 1 was informed by the research discussed in Section 2.5 highlighting the negative impact of inter-group bias on in-
dividuals from a variety of groups (Fries-Britt & Turner, 2001; Jacoby-Senghor et al., 2016; Kellow & Jones, 2008; Stauffer & Buckley, 2005). Therefore, Research Question 1.0 was investigated:

**RQ1.0: How do computational approaches, that claim to uncover inter-group bias in non-educational texts, perform in text-based online learning materials?**

To answer that question, literature for existing approaches aimed at uncovering inter-group bias in texts was reviewed (Section 2.6). Afterwards, two promising approaches to identify bias in online texts (i.e., Gaucher et al. (2011) and Pryzant et al. (2020)) were selected based on the literature review (Section 2.6.2). Then, those approaches were applied to the context of an online learning platform and their outputs were assessed (see Chapter 4 for details). While the theoretical portion of Study 1 (i.e., identifying relevant approaches in the literature) was conducted during the literature review (i.e., in Section 2.6.2), the practical portion of Study 1 was organised as one research cycle of Analysis and Exploration, Design and Construction, and Evaluation and Reflection (i.e., Research Cycle 1).

**Research Cycle 1 (RC1)**

In essence, RC1 comprised the initial exploration regarding the underlying problem behind this thesis: inter-group bias in online learning. As indicated in the Literature Review (Chapter 2), it was unclear to what extent existing mechanisms to uncover inter-group bias in non-educational contexts would be suitable for identifying such biases in text-based learning materials from online learning platforms. Therefore, RC1 culminated in the need for looking at and evaluating existing approaches as a means to identify their limitations within learning settings. Accordingly, two evaluated strategies were adopted: (i) randomness check, which comprised testing the classification of each approach (i.e., potential biases) against arbitrary classification – a common practice to determine whether an algorithm is better than just guessing randomly (James et al., 2013); and (ii) human evaluation, which focused on com-
paring student opinions regarding what they would consider (not) biased against the potential biases suggested by each approach. In essence, RC1 focused on answering the following secondary research questions:

- **RQ1.1:** How do the approaches’ outputs compare to random classification?
- **RQ1.2:** How do the approaches’ outputs compare to student classifications?

Table 3.3 shows the key methodological aspects of Study 1 and, therefore, RC1. More details about the methodological aspects of this study are provided in the respective chapter (Chapter 4).

### Table 3.3: Overview of Study 1

<table>
<thead>
<tr>
<th>Study</th>
<th>Study 1 (Chapter 4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Core RQ</td>
<td>RQ1.0: How do computational approaches, that claim to uncover inter-group bias in non-educational texts, perform in text-based online learning materials?</td>
</tr>
<tr>
<td>Research cycle</td>
<td>RC1</td>
</tr>
<tr>
<td>Analysis &amp; Exploration</td>
<td></td>
</tr>
</tbody>
</table>
| **What is known?** | - Bias can harm students  
- Approaches to identify bias were proposed for non-educational settings |
| **What is not known?** | The extent to which such approaches would be suitable for educational settings |
| Sub RQs | RQ1.1  
RQ1.2 |
| Focus | To understand how existing approaches perform in educational settings |
| Design & Construction | |
| Data | - 2024 sentences from learning materials of the OU VLE  
- 20 HE students from the UK and US |
| **Implementation** | 1. The download of learning materials and sentence extraction  
2. Sentences checked for potential biases using the approaches previously selected |
| **Evaluation** | 3. Selection of a subset of 30 sentences  
4. HE students were asked to assess the 30 sentences for potential biases  
5. Students’ marks compared to approaches’ |
| Key methods | Data Acquisition  
- Web data extraction  
- Survey |
| Data Analysis | - Descriptive statistics  
- Inferential statistics |
| Evaluation & Reflection | Key outputs | - Evidence of how existing approaches perform in educational settings  
- Insights about the data source and recruiting platform  
- Design implications for the next research cycle |

For more details about Study 1, please refer to Chapter 4.

### 3.4.3 Study 2 (see Chapter 5)

Study 2 was informed by the results of the previous study and aimed at understanding how contextual elements affect the identification of inter-group bias in learning
materials and how such biases were perceived by students. Accordingly, Study 2 was guided by the following research question:

*RQ2.0: How does bias against certain ethnic groups manifest in text-based online learning materials?*

To answer that question, Study 2 was conducted in two research cycles (RC2 and RC3). RC2 looked at how to design an intervention to capture the perspectives of particular ethnic groups about ethnic bias in text-based learning materials (e.g., which dataset is suitable, how many participants per group, etc.). Afterwards, RC3 focused on which attributes of learning materials (e.g., discipline, title) could provide some context to support the identification of ethnic bias in text-based learning materials and how individuals from different ethnic groups perceive ethnic biases.

**Research Cycle 2 (RC2)**

RC2 was centred on methodological aspects of Study 2 rather than answering its core research question. Accordingly, RC3 was designed as a small intervention aiming at uncovering potential unforeseen issues before implementing a large-scale intervention. RC2 was guided by the following sub-question:

*RQ2.1: How do we operationalise contextual and cultural aspects to support the identification of ethnic bias in text-based online learning materials?*

The design of RC2 comprised selecting a data source and participants, getting the data, planning an online intervention, and defining the data analysis procedure. Open educational resources were identified as promising data sources, as their open licences would allow their extraction for research purposes (see Section 5.2.2). In addition, the literature review suggested their metadata (e.g., the material’s title, type, and discipline) as potential means to inform their context. The online platform Prolific was identified as a promising channel for recruiting participants, as it allows pre-screening of participants based on their demographics before they start

the study. Once the data and participants were defined, an online task was created. The task comprised a labelling task, where participants assessed a set of sentences for potential ethnic bias.

**Research Cycle 3 (RC3)**

RC3 looked at how students perceive ethnic bias, in order to establish a baseline upon which LA models could be built. In addition, it set out to highlight potential causes of ethnic bias in text-based OERs. This research cycle was mostly motivated by the findings of RC1 that suggested student perspectives of ethnic bias have not gained much attention from the research community. In addition, the design of RC3 was informed by RC2, which provided insights about this design of RC3. In particular, RC3 focused on the following sub-questions:

- **RQ2.2**: How do students from certain ethnic groups perceive ethnic biases in text-based online learning materials?

- **RQ2.3**: To what extent do contextual attributes extracted from text-based online learning materials support the identification of potential ethnic bias?

- **RQ2.4**: What makes text-based online learning materials potentially biased or not biased according to students?

The design of RC3 comprised an online task similar to the ones in the previous research cycles. In this stage, a larger dataset containing 325 sentences with context attributes was split into 22 batches (15 sentences each) and made available online for ethnic bias assessment. For the assessment, 132 students were recruited and divided into groups of 6 students per batch. They were asked to label each sentence in terms of ethnic bias and to indicate the reason for their responses. Once the data were collected, two types of analysis were conducted: a quantitative analysis aiming at identifying which sentences were perceived as biased by students from different ethnic groups; and a qualitative analysis looking at potential reasons for ethnic bias.
in learning materials.

Table 3.4 provides an overview of Study 2. More details about the methodological aspects of this study are provided in the respective chapter (Chapter 5).
Table 3.4: Overview of Study 2

<table>
<thead>
<tr>
<th>Study: Study 2 (Chapter 5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Core RQ: RQ2.0: How does bias against certain ethnic groups manifest in text-based online learning materials?</td>
</tr>
<tr>
<td>Research cycles: RC2  RC3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Analysis &amp; Exploration</th>
<th>What is known?</th>
<th>What is not known?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>- Approaches to identify bias in non-educational settings do not seem effective for learning materials</td>
<td>- How to capture what might be considered biased to certain ethnic groups and, at the same time, take the context of learning materials into consideration</td>
</tr>
<tr>
<td></td>
<td>- What is considered biased might depend on context and culture</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Classification models are strongly bound to the data used to train them but labelled data about ethnic bias in learning texts seem scarce</td>
<td>- What makes learning materials potentially biased</td>
</tr>
<tr>
<td></td>
<td>- The literature suggests that inter-group bias might depend on context and culture, but evidence about that assumption for ethnic bias in learning materials is limited</td>
<td>- The actual relevance of context when students are assessing ethnic bias</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sub RQs</th>
<th>RQ2.1</th>
<th>RQ2.2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>RQ2.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RQ2.4</td>
</tr>
</tbody>
</table>

| Focus | To design an intervention to capture student perceptions of what might be biased (considering their ethnic group and the context of learning materials) | To uncover potential ethnic biases in texts from learning materials and, at the same time, account for the context of such materials and the perspectives of students from ethnic minority groups. |

<table>
<thead>
<tr>
<th>Design &amp; Construction</th>
<th>Data</th>
<th>- 20 sentences from Open Educational Resources</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>- 61 HE students from the UK and US</td>
</tr>
<tr>
<td></td>
<td>Key procedures</td>
<td>- 325 sentences from Open Educational Resources</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- 132 HE students from the UK and US</td>
</tr>
<tr>
<td>Data</td>
<td></td>
<td>1. Defining the data source and selecting the sentences</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2. Designing and implementing an online task</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3. Designing an evaluation approach</td>
</tr>
<tr>
<td>Key procedures</td>
<td></td>
<td>4. Participants recruitment (data collection)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5. Data analysis</td>
</tr>
<tr>
<td></td>
<td>Key methods</td>
<td>1. Scaling up the task from RC2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2. Defining the data analysis pipeline</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3. Recruiting participants (data collection)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4. Data analysis</td>
</tr>
<tr>
<td>Data Acquisition</td>
<td></td>
<td>Data Acquisition</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Web data extraction (extracted during RC2)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Survey</td>
</tr>
<tr>
<td>Data Analysis</td>
<td></td>
<td>Data Analysis</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Descriptive statistics</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Inferential statistics</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Thematic Analysis</td>
</tr>
<tr>
<td>Evaluation &amp; Reflection</td>
<td>Key outputs</td>
<td>- Evidence of how ethnic bias manifest in learning texts</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Dataset with 345 contextualised sentences from OERs</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Dataset with 345 labelled sentences from OERs</td>
</tr>
</tbody>
</table>

For more details about Study 2, please refer to Chapter 5.
**3.4.4 Study 3 (see Chapter 6)**

Study 3 focused on utilising established Learning Analytics models to assess the dataset generated in Study 2, aiming to evaluate their effectiveness in classifying ethnic bias. In particular, this study aimed to address the research question:

*RQ3.0: To what extent can Learning Analytics assist in the identification of ethnic bias in text-based online learning materials?*

Accordingly, commonly LA models were applied to the Study 2 dataset, with their performance assessed in relation to identifying perceived ethnic bias within textual OERs. As previously mentioned, Study 3 encompassed two research cycles: (i) RC4, which targeted the selection of pertinent features for input in machine learning algorithms; and (ii) RC5, which sought to evaluate the efficacy of widely used classification methods in the context of ethnic bias identification within learning texts.

**Research Cycle 4 (RC4)**

RC4 was informed by both literature and findings from the previous cycle. For instance, a literature review was conducted aiming at highlighting textual features commonly used in text classification that could be extracted from the dataset built during this research project. In particular, RC4 focused on the following sub-question:

*RQ3.1: Which features might support the identification of ethnic bias in text-based online learning materials?*

The design of RC4 comprised creating new features from the existing dataset. While a detailed rationale is provided in Chapter 6 (Section 6.2.1), the feature expansion process focused on five aspects informed by Study 2 and the literature: (i) psycholinguistic markers, which are signs that show the author’s mental state during writing (Sboev et al., 2015); (ii) linguistic abstraction – the use of different words
(e.g., verbs and adjectives) to portray a person or behaviour, varying from specific to general levels of detail (Tincher et al., 2016); (iii) language valence, which describes a component of emotion that relates to how positive or negative a stimulus is perceived (Osgood et al., 1957); (iv) ethnic group mentions, which consisted of explicit words referring to an ethnic group (e.g., Asian teacher, Black student, etc.); and (v) contextual attributes – characteristics of online learning materials that could indicate the context from which they were extracted (e.g., discipline).

Accordingly, once features were expanded according to those aspects, relevant features were selected based on the filter method, which involves selecting features based on statistical measures (e.g., correlations, chi-squared test, etc.) (Guyon & Elisseeff, 2003). This approach was chosen because it gives the possibility of quickly identifying the most relevant features without involving complex techniques (e.g., machine learning model), which is useful for reducing computational complexity (Chandrashekar & Sahin, 2014). According, RC4 used a correlation approach in order to identify features which were potentially related to ethnic bias (see Section 6.2.2 for detailed steps). Once relevant features were identified, they could be used to train a range of classification models as highlighted below.

**Research Cycle 5 (RC5)**

RC5 comprised using the features from the previous research cycle for training different classifiers against the identification of ethnic bias. In addition, this research cycle was informed by the literature in terms of commonly used machine learning algorithms for classification tasks. In particular, this research cycle focused on the question:

*RQ3.2: Which classification approaches might be suitable for identifying ethnic bias in text-based online learning materials?*

The design of RC5 comprised identifying machine learning models and defining evaluation metrics. For instance, the following models were tested: (i) Logistic Re-
CHAPTER 3. METHODOLOGY 3.4. RESEARCH DESIGN

gression (LG); (ii) Decision Tree (DT); (iii) Naive Bayes (NB); (iv) Support Vector Machines (SVMs); (v) K-Nearest Neighbours (KNN); (vi) Random Forest (RF); and (vii) Extreme Gradient Boosting (XGB, also known as XGBoost). The construction consisted of formatting the dataset according to each approach, implementing and executing the models, and computing performance metrics.

Table 3.5 provides an overview of Study 3. More details about the methodological aspects of this study are provided in the respective chapter (Chapter 6).
Table 3.5: Overview of Study 3

<table>
<thead>
<tr>
<th>Study:</th>
<th>Study 3 (Chapter 6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Core RQ:</td>
<td>RQ3.0: To what extent can Learning Analytics assist in the identification of ethnic bias in text-based online learning materials?</td>
</tr>
<tr>
<td>Research cycles:</td>
<td>RC4  RC5</td>
</tr>
<tr>
<td>Analysis &amp; Exploration</td>
<td></td>
</tr>
<tr>
<td>What is known?</td>
<td>- Different aspects of language have already been used to analyse and explore text content</td>
</tr>
<tr>
<td>What is not known?</td>
<td>- Which aspects of language (features) could be relevant for identifying ethnic bias in text</td>
</tr>
<tr>
<td>Sub RQs</td>
<td>RQ3.1  RQ3.2</td>
</tr>
<tr>
<td>Focus</td>
<td>To explore potential features that could be used for classifying potential ethnic bias in learning texts</td>
</tr>
<tr>
<td>Design &amp; Construction</td>
<td></td>
</tr>
<tr>
<td>Data</td>
<td>345 labelled sentences from OERs (from Study 2)  345 labelled sentences from OERs (from Study 2)</td>
</tr>
<tr>
<td>Key methods</td>
<td>- Filter method for feature selection  - Inferential statistics</td>
</tr>
<tr>
<td>Evaluation &amp; Reflection</td>
<td></td>
</tr>
<tr>
<td>Key outputs</td>
<td>- Evidence of which aspects of learning materials might be associated with potential ethnic bias  - Potential features for automating the classification of ethnic bias in text</td>
</tr>
</tbody>
</table>

For more details about Study 3, please refer to Chapter 6.
3.5 Data & Ethics

This section covers the steps taken to ensure this project was ethical and in accordance with data protection regulations.

3.5.1 Ethics review

To make sure this research was in compliance with ethics standards, all methodological approaches and research interventions that comprised sensitive data were submitted for the appreciation of the local ethics review board (i.e., The Open University Human Research Ethics Committee - HREC). As the studies were iterative and sequence-dependent (i.e., the design of the later studies depended on the results of the previous ones), submissions to the HREC committee were made on demand (i.e., according to the needs of each study). Accordingly, an initial application focused on the data and procedures of Study 1 (Chapter 4). Then, an amendment request was made to cover the data and procedures of studies 2 and 3 (chapters 5-6). This rationale was adopted to ensure that only the necessary data were collected, which is a requirement of the General Data Protection Regulation (GDPR)\(^5\). The HREC approval letter is available in Appendix A.2.

3.5.2 Data protection

All data used in this project were selected according to the aims of each study and comprised text-based learning materials extracted from online learning platforms, and survey data from Higher Education students from the UK and the US. In particular, Study 1 used audio-visual transcriptions from learning materials extracted from the OU Virtual Learning Environment (VLE), and survey data from the Amazon Mechanical Turk\(^6\) (AMT) platform. Similarly, the learning materials used in

\(^5\)https://gdpr-info.eu/
\(^6\)https://www.mturk.com/
Study 2 and Study 3 were extracted from the OER Commons Platform\(^7\), and the survey data were collected via the Prolific platform\(^8\).

While more details about the rationale for those data are available in the respective “design” sections of each study (i.e., Section 4.2.2, Section 5.2.2, and Section 5.3.2), the overall rationale can be summarised as follows:

- The need for collecting text-based learning materials assumed that there were potential biases in online learning texts (see Section 1.2).

- The need for collecting students’ demographics (i.e., ethnic group, country of origin, and level of education) was informed by one primary objective of this project: to understand how students from different groups perceive ethnic bias in online learning texts (see Section 1.3).

- The collection of contextual attributes of text-based learning materials (i.e., title, type, discipline, and excerpt) was informed by the need for providing context to students when they were assessing potential ethnic biases in those materials (see Section 5.2.2). In addition, those attributes were also needed for training the machine learning models implemented in Study 3 (see Section 6.3.2).

Furthermore, all potential risks to individuals were carefully considered. For instance, the potential risks of this research involved exposing students’ and/or tutors’ information (i.e., some learning materials contained personal information). To mitigate those risks, the following strategies were followed, (i) all data that could potentially identify individuals were anonymized (e.g., student’s and courses’ identifiers, tutor’s name); (ii) findings were reported in terms of groups (e.g., gender, ethnicity, knowledge areas) instead of individuals; (iii) potential sensitive data were stored in password-protected computer drives, and adequate security technologies were used (e.g., antivirus, secure data transfer protocols, encrypted storage). In ad-

\(^7\)https://www.oercommons.org/
\(^8\)https://www.prolific.co/
dition, all personal data were treated as being strictly confidential and only this PhD researcher and his supervisors had access to them.

In addition to seeking HREC approval and following the aforementioned steps, the present student also completed two relevant modules about information security and GDPR: (i) Information Security Awareness module (see Appendix B.2), and (ii) GDPR (see Appendix B.3).

### 3.6 Chapter summary

This chapter presented the underlying methodology of this thesis. First, the research philosophy was introduced, and its core aspects were indicated in terms of ontology, epistemology, and axiology. Second, it was shown how those aspects fit within a pragmatic paradigm, and how they informed the methodological choice: Design-Based Research. Third, the fundamentals of DBR were summarised and the rationale for using it on this project was highlighted. Once a methodological base was established, the research design was presented. The present project is organised into three studies and six research cycles: Study 1 (one research cycle), Study 2 (two research cycles), and Study 3 (two research cycles). The latter part of this chapter described the ethics and data protection aspects related to this research.

More details about each study are provided in the subsequent chapters, beginning with Study 1.
4 | Study 1

4.1 Chapter overview

This chapter covers Study 1 of this PhD project, i.e., an initial study that was organised into one research cycle (RC1) as indicated in the previous chapter (Section 3.4). Study 1 was motivated by limited evidence regarding the suitability of existing bias-classification approaches for classifying inter-group bias in text-based learning materials (see Section 1.2). In addition, Study 1 was informed by the research discussed in Section 2.5 highlighting the negative impact of inter-group bias on individuals from a variety of groups (Fries-Britt & Turner, 2001; Jacoby-Senghor et al., 2016; Kellow & Jones, 2008; Stauffer & Buckley, 2005). Accordingly, this chapter is organised as follows: Section 4.2.1 presents the analysis and exploration stages of RC1, which indicates the problem concerned in this step. Section 4.2.2 presents the research design, which includes the steps undertaken to uncover existing computational approaches for bias identification, the methods used for implementing them, and the procedures followed to evaluate them in the present context. Section 4.2.3 presents the results, the discussion, and indicates core implications for the next research cycle (RC2).

4.2 Research Cycle 1

4.2.1 RC1: Analysis and Exploration

As indicated in Chapter 2, online learning technologies have incorporated human prejudices and biases (de Souza & Perry, 2021; Zhang, 2015). It was also shown

\footnote{The research conducted during Study 1 is being prepared for publication. Preliminary details:}

that, when not addressed, those biases can negatively affect students of particular social groups in different ways (Albuquerque et al., 2017; Berg et al., 2018; Greenwald & Krieger, 2006; Richardson et al., 2020), for example, by triggering psychological mechanisms like anxiety that ultimately affect their academic performance (Pennington et al., 2016). Despite those findings highlighting the existence of biases in online learning and their consequences to students, mitigating its negative impact is challenging. In particular, the large amounts of online educational data make it impractical to uncover such biases by hand, for example, as done in corpus linguistics studies described in Section 2.6.1. Indeed, a recent report from Research and Markets (2022) estimated growth in the digital education market size from USD 11.5 billion in 2021 to USD 46.7 billion by 2026, which directly impacts the amount of educational data available online.

Chapter 2 (Section 2.6) also indicated that researchers have already proposed various means to identify inter-group bias (see Section 2.4 for formal definition), which range from manual approaches (e.g., questionnaires, research experiments) to computational approaches (e.g., machine learning algorithms). While manual approaches seem unsuitable for identifying bias in large datasets, it seems there is limited evidence regarding the suitability of computational strategies for identifying bias in educational settings (see Section 2.6 for more details). Accordingly, this research cycle aimed at exploring whether potential inter-group biases from text-based online learning materials could be effectively identified with generic computational approaches. In particular, RC1 focused on answering the following research question:

**RQ1.0: How do computational approaches, that claim to uncover inter-group bias in non-educational texts, perform in text-based online learning materials?**

To address this research question, the study was divided into two parts: (i) implementing the computational approaches identified in the literature (Section 2.6.2) in an educational context, and (ii) evaluating the performance of these approaches.
The operationalisation of these parts is described in the following section.

4.2.2 RC1: Design and Construction

This stage focused on designing a way to implement existing approaches aiming at identifying bias in text in an educational setting so that potential limitations regarding the use of such approaches in learning settings could be highlighted. In particular, the whole process took place in four main steps (more details are provided in the subsequent sections):

1. **Approach selection.** Two computational approaches aiming at identifying bias in text were selected for implementation. Those approaches were the ones found during the literature review and were presented in Section 2.6.2. More details about each one are also presented in this chapter.

2. **Data selection.** A dataset containing 3200 sentences sampled from 91 courses of an online learning platform was also selected for testing the approaches indicated above.

3. **Implementation.** The two approaches were applied to the 3200 sentences. This step produced a list of sentences that were potentially biased (i.e., sentences that each approach classified as biased).

4. **Evaluation.** The evaluation comprised two parts:

   • **Part 1: Randomness check.** The potential biases (as indicated by each approach) were checked for potential randomness. This would suggest to what extent the results (potential biases) of each approach were arbitrary.

   • **Part 2: Human evaluation.** Higher Education\(^2\) (HE) students were also recruited and asked to label a sample of such sentences as *biased* or *not biased*.

\(^2\)HE was chosen due to the prevalence of research on bias in online learning in this context (Section 2.5).
biased. This step aimed at comparing student opinions regarding what they would consider (not) biased against the potential biases suggested by each approach.

The four steps are also illustrated in Figure 4.1 and detailed after the next section (data context and rationale). Before selecting the approaches and extracting the data, the context for this research will next be defined.

![Figure 4.1: Overview of RC1](image)

**Data context and rationale**

The context for this research cycle was The Open University UK (OU) Virtual Learning Environment (VLE). In particular, the data used in this research cycle comprised transcriptions of online learning materials extracted from online videos. This content was selected because the presence of inter-group bias in such an environment could potentially impact a huge group of individuals. The OU is the largest university in the United Kingdom in terms of undergraduate students and one of the largest universities in Europe with more than 170k enrolments (University, 2019).

In addition, transcriptions from online videos were used because it was expected that they would have a stronger probability of containing potential biases when compared to other online learning materials. As identified by Rienties and Toetenel
(2016b), the quality process of producing online materials at the OU is very intensive, with a range of people checking, editing, and refining content before it goes online. Given such an intensive review process, it was expected that limited inter-group bias would remain in edited texts. In contrast, one notable exception to this multi-author review process is the production of online videos, which are often produced by a single author and uploaded at the end of a production cycle. As these videos are produced and reviewed by a limited number of people, there was a stronger probability that inter-group biases could be included. Therefore, only materials from OU online videos were considered for this initial research cycle.

Furthermore, as part of the accessibility requirements, all OU videos contain transcriptions. That would make such materials suitable for this research cycle to explore whether any inter-group bias could be present (or not) in such text-based materials. The transcriptions belonged to a range of disciplines including Arts, Education, Health, Computing, Natural Sciences, Business, Law, and Social Sciences. Figure 4.2 illustrates those transcriptions through two examples.

**Example 1**

*Welcome message from the module team*

View site website

Hello and welcome to this introductory video. My name is ..., I'm a senior lecturer in ..., and I'm here at the OU. We want to use this opportunity to give you some insights into what this module is all about. I'm joined today by two colleagues, ..., and one of our graduating students, ...

So ..., can I start with you? Can you briefly describe what the module is all about and what students will benefit from it?

*Example 2*

*Video 3.1 Doing titrations.*

View on website

NARRATOR: A titration is a classical analytical technique where an unknown quantity of a particular substance is determined by adding it to a standard solution. That's one that we know the concentration of to a high degree of accuracy. We can calculate the concentration of the unknown by referring to a balanced chemical equation for the reaction.

Modern analytical laboratories have autotitrators which largely do the job for us, and allow a large throughput of samples. But if you are asked to carry out a titration manually, on the bench, using standard laboratory glassware, then this is the apparatus you'll need.

We're going to demonstrate an acid-base titration. Our unknown is a solution of sodium hydroxide, and our standard is a point one mole per litre solution of hydrochloric acid. We prepare this by diluting a commercially available standard. The label tells us we need to dilute to 500 millilitres.

Notice we wash the bottle out with deionised water. And make up

**Figure 4.2: Excerpts from two OU video transcriptions**

Once the research context was defined, the key steps of this research cycle were
initiated. As indicated previously, those steps comprised approach selection, data selection, implementation, and evaluation. More details are provided as follows.

Approach Selection

While the two approaches implemented in this research cycle (RC1) were already presented in Chapter 2 (Section 2.6.2), further key aspects of how they work are presented below. In addition, extensive details regarding their original designs and implementations are available within the authors’ respective publications: Gaucher et al. (2011) and Pryzant et al. (2020).

- Approach 1: “Evidence That Gendered Wording in Job Advertisements Exists and Sustains Gender Inequality” by Gaucher et al. (2011)

This approach was selected for implementation because it is directly related to the scope of this research cycle, i.e., it allows the identification of potential inter-group bias in text-based content. In addition, it has informed a number of studies examining gender bias in a variety of contexts, such as workplace (Heilman, 2012; Sczesny et al., 2016) and education (González-Pérez et al., 2020), which demonstrates its versatility and potential for identifying biases in the context of this research. Furthermore, Gaucher et al. (2011) conducted five empirical studies that focused on gender bias in job advertisements. They found statistically significant results suggesting the presence of gender bias in job advertisements and showed that such biases could affect individuals’ perspectives of belonging regarding gender-dominated occupations.

The approach used by Gaucher et al. (2011) comprised lists of masculine and feminine words to analyse job advertisements. These lists were derived from published lists of agentic and communal words, as well as masculine and feminine trait words, which were established in earlier literature (e.g., Bartz and Lydon (2004), Rudman and Kilianski (2000), and Hoffman and Hurst (1990)). Combining these pre-established lists, they coded the adverts to determine the percentage of masculine
and feminine words present in each advertisement, which allowed them to systematically identify potential gender bias in the language used in such contexts.

In this research cycle, the two resulting lists of gender-dominant words utilised by Gaucher et al. (2011) were used to flag potential gender bias in sentences extracted from transcriptions of online learning materials (more details about the data are presented in the following section). For instance, sentences containing a predominant number of words from a gender-dominant list would be considered biased towards that gender. For example, as shown in Section 2.6.2, the sentence “You are confident and ambitious in these competencies” would be considered biased because it contains more words from the male-dominant list (i.e., ‘confident’, ‘ambitious’, and ‘competencies’) than the female-dominant list (none). An excerpt of those lists of words considered gender-dominant is also available in Section 2.6.2, Table 2.1.

- Approach 2: “Automatically Neutralizing Subjective Bias in Text” by Pryzant et al. (2020)

This approach was selected for implementation as it has been demonstrated by Pryzant et al. (2020) to identify biases associated with demographic variables such as ethnicity and gender, potentially encompassing inter-group biases. The biases that the approach detects are referred to as subjective bias, which is characterised using language that presents a text as biased based on opinions, emotions, or personal preferences (Pryzant et al., 2020). Section 2.6.2, Table 2.2 presents some instances of bias identified by Pryzant et al. (2020).

In contrast to the previous approach, Pryzant et al. (2020) used natural language processing (NLP) techniques to train two models that predict bias in sentences and suggest an equivalent neutral sentence. In particular, the underlying models were trained using the Bidirectional Encoder Representations from Transformers (BERT) approach, an NLP method considered “state-of-the-art” in several tasks by the time of its publication (Devlin et al., 2019). The models were trained based on 180,000 sentence pairs extracted from Wikipedia’s history of edits. Each pair consisted of an
original sentence considered by Wikipedia’s editors as “biased”, and a neutralised version considered by such editors as “neutral”. Those judgements were based on Wikipedia’s neutrality principle (Wikimedia Foundation, 2020) – a principle indicating that Wikipedia’s content should be neutral and free of biases:

All encyclopaedic content on Wikipedia must be written from a neutral point of view (NPOV), which means representing fairly, proportionately, and, as far as possible, without editorial bias, all the significant views that have been published by reliable sources on a topic. (Wikimedia Foundation, 2020, para. 2)

In essence, Approach 2 comprised two algorithms: one to automatically label sentences as biased or not biased (reported accuracy = 0.759); and another one (reported accuracy = 0.745) to neutralise the sentences flagged as ‘biased’ (i.e., once a sentence was labelled “biased”, the second algorithm would edit such a sentence to make it “neutral” or “unbiased”). While both algorithms seemed promising in the context of Wikipedia pages, only the first part (the one regarding bias identification) was implemented in this research cycle as neutralising bias was not in the scope of this PhD Project.

To implement both approaches in the OU VLE context, first, the data from such context needed to be acquired and pre-processed. More details about the data selection and pre-processing are provided as follows.

**Data selection**

To select the data for this research cycle, a set of data selection criteria were defined. For instance, only courses taught in English, and that had a transcription available were included. Those criteria were defined in order to ensure that only text-based and content written in English would be collected, which was based on the scope delimitation of this PhD Project (i.e., bias in text, as shown in Section 1.1). In addition, to ensure the courses had already been completed and, at the same time, to guaran-
tee they were as recent as possible, only transcriptions from courses that started in October 2018 were included. Such a criterion would guarantee that courses would have finished at the time RC1 started (i.e., in October 2019). Furthermore, older course transcriptions could potentially make the findings of RC1 not applicable to current online learning platforms as such settings have been constantly evolving.

Initially, a total of 300 courses were screened, but 181 were excluded for not having transcriptions associated with them. From the remaining 119 courses, 22 were from language subjects and were excluded for having non-English content. In addition to those exclusions, another 6 courses were removed to avoid the authors being identified (too few courses per discipline). This process resulted in transcriptions from 91 courses containing a total of 3200 sentences, from which several of them (n = 1176) was ultimately excluded for various reasons. For example, when using the Pryzant approach, an initial step comprised tagging parts of speech of each sentence (i.e., labelling words as nouns, adjectives, articles, etc.); when a given sentence included words that were not encoded in the tagging model, the sentence was skipped. In essence, the final dataset considered for analysis contained 2024 sentences.

Finally, those transcriptions were downloaded from the OU VLE\(^3\), and the sentences were pre-processed.

Pre-processing steps were required to avoid data inconsistencies and to format raw data into standardised input so that they could be compatible with the selected approaches. In addition, these steps aimed to remove irrelevant attributes in order to reduce the dataset complexity. Based on those aspects, the steps below were undertaken before the implementation.

1. **Data extraction.** This included the collection of the course discipline, its identifier and transcription, based on the criteria mentioned above.

\(^3\)https://learn2.open.ac.uk/
2. **Data sanitising.** Document markers that were not needed for inferring potential biases were removed from each transcription (e.g., formatting tags, markers of sessions and headings).

3. **Sentence extraction.** Each transcription was split into sentences.

4. **Sample selection.** To prevent selection bias and reduce the data processing time, samples were randomly selected from each discipline to provide a sample size of 400 sentences per discipline. That number was used in order to achieve a confidence level of 95% while keeping the margin of error below 10%.

5. **Data anonymisation.** This step comprised removing individuals’ names from texts. This was a requirement to comply with ethics standards, i.e., for preventing the identification of tutors or students in the results. In order to complete this step, the Named Entity Recognition approach from Stanford University (Finkel et al., 2005) was used. The process involves identifying and classifying named entities (e.g., person names, organisations, locations, etc.) automatically, which is more convenient when dealing with large datasets. In addition, all sentences presented along with this chapter were manually checked to prevent any potential misses. Therefore, the future publication of this research cycle would not disclose any potential personal information present in the reported data.

6. **Data format.** Finally, the extracted data were formatted to meet the input requirements of both Pryszt’s and Gaucher’s approaches.

At the end of the pre-processing stage, a further number of sentences \((n = 1176)\) were ultimately excluded for various reasons. For example, the data format stage comprised tagging parts of speech of each sentence (i.e., labelling words as nouns, adjectives, articles, etc.). When a given sentence included words that were not encoded in the tagging model, such a sentence was skipped. Although tagging
parts of speech was a pre-processing requirement specific to the Pryzant approach, the same sentences were intentionally prevented from being processed by Gaucher approach in order to benefit from the comparison between the two approaches. In essence, the final dataset used for implementing the two approaches comprised a total of 2024 sentences after the exclusions performed during the data selection and pre-processing steps. Below, further details about the implementation are provided.

**Implementation**

Once the data were pre-processed and ready for use, the implementation process started. It involved (i) calculating the occurrence of gender-dominant words using the dataset from Gaucher et al. (2011); and (ii) applying the algorithm provided by Pryzant et al. (2020). Given that Gaucher’s approach did not require high processing performance, a personal computer (HP EliteBook 820 G3) was used to identify potential gender-dominant words in the sentences. Pryzant’s strategy was available online\(^4\) and was coded in the Python language. Since their technique demanded high computational processing power, the data were processed in a remote machine provided by Amazon Web Services (AWS). An Amazon Elastic Compute Cloud (EC2) of type p3.2xlarge was allocated, which comes with 1x GPU Tesla V100, 16GB GPU Memory, 8x vCPUs, and 61GB Memory.

**Evaluation**

The evaluation comprised two parts, first descriptive statistics were used and compared the classifications of each approach to random chance. This strategy was chosen based on the assumption that each approach would classify the sentences differently from an arbitrary classifier, i.e., the number of potential biases within the dataset would differ from 50\%. The results of this first part are reported in terms of discipline to unveil a better perspective regarding the context (Arts, Education, Health, Computing, Natural Sciences, Business, Law, and Social Sciences).

\(^4\)https://github.com/rpryzant/neutralizing-bias
In the second part, Higher Education students were asked to evaluate a subset of the data in terms of inter-group bias aiming at understanding how human perspectives about what is potentially biased would compare to the approaches’ classifications (see Section 4.2.3). More details are provided below.

**Part 1: checking the results for random chance**

This part focused on answering the question:

\textit{RQ1.1: How do the approaches’ outputs compare to random classification?}

Accordingly, in order to provide insights into how each approach inferred potential biases within the online learning texts, descriptive statistics were computed, and their classifications were checked for randomness. In other words, relative frequencies and proportions were calculated, and the null hypothesis (\(H_{10}\): the classifications result from random chance) was tested. Given both approaches tried to classify the sentences into biased or not biased, it was assumed that a random classification would classify 50% of sentences as biased \([P(\text{BIAS}) = .5]\) and 50% as not biased \([P(\neg\text{BIAS}) = .5]\). It was also assumed the classification of each sentence was independent, and that the only source of variation was random and binomial.

**Part 2: human evaluation**

This part aimed to gather human insights regarding the performance of Pryzant and Gaucher approaches applied to online learning texts. In other words, Part 2 focused on answering the question:

\textit{RQ1.2: How do the approaches’ outputs compare to student classifications?}

In order to execute this step, a crowd-sourcing strategy was used for recruiting participants. Crowd-sourcing gives the possibility of recruiting participants with a diverse mindset and promoting sample representativity (Palmer & Strickland, 2016). Accordingly, an online task was set up on the Amazon Mechanical Turk (AMT) platform, which can to provide large amounts of data quickly and inexpensively
(Chandler et al., 2014). AMT was selected based on the researcher’s familiarity with it (i.e., the present student had previously used AMT in other projects), which would reduce the implementation time. In addition, AMT has predefined tasks (“templates”) tailored for labelling data, which aligned with the aims of this research cycle.

For the online task, a subset of sentences (n = 30) was randomly extracted from the list of sentences processed by both Pryzant’s and Gaucher’s approaches (n = 2024). Then, for each of the 30 sentences, an assignment was created on the platform asking the question “Is this sentence biased?” The possible answers were “yes”, “no”, and “can’t tell”. In addition, the following definition of bias was given: “A sentence is biased when it is inclined towards a particular group of different gender, race, economic background, etc.”, and two examples were also provided:

Example 1:

- Biased: “A nurse is trained to understand her patients”
- Not biased: “Nurses are trained to understand their patients”

Example 2:

- Biased: “Mankind has developed amazing technologies”
- Not biased: “Humanity has developed amazing technologies”

In Example 1, the word “her” makes the term “nurse” suddenly associated with females, which makes the sentence biased. In the second example, the sentence “Mankind has developed amazing technologies” is biased because the word “mankind”, in Middle English, means “male persons”, which may exclude female individuals. The second sentence on each example suggested a way to prevent the respective biases.

Table 4.1 provides the sentences used in this task (i.e., the sentences in which stu-
Table 4.1: Sample of sentences submitted to participants to assess potential biases.

<table>
<thead>
<tr>
<th>id</th>
<th>sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>and I’ll maybe catch up with you again at another time.</td>
</tr>
<tr>
<td>S2</td>
<td>and I was like, ok, let’s do it, whatever we’ve got to do.</td>
</tr>
<tr>
<td>S3</td>
<td>and it’s probably the only way that leads out of the crisis.</td>
</tr>
<tr>
<td>S4</td>
<td>and what’s your reaction to the court’s decision today?</td>
</tr>
<tr>
<td>S5</td>
<td>and you don’t argue and you don’t shout, ok?</td>
</tr>
<tr>
<td>S6</td>
<td>he’s not hurt me.</td>
</tr>
<tr>
<td>S7</td>
<td>I haven’t put the solar wind in yet, but I will shortly.</td>
</tr>
<tr>
<td>S8</td>
<td>I’ll stop for just a second.</td>
</tr>
<tr>
<td>S9</td>
<td>I mean who am I to I don’t think it matters.</td>
</tr>
<tr>
<td>S10</td>
<td>I think it’s a rubbish idea.</td>
</tr>
<tr>
<td>S11</td>
<td>I’ve got two at 26.</td>
</tr>
<tr>
<td>S12</td>
<td>it’s a community nursery in a deprived area.</td>
</tr>
<tr>
<td>S13</td>
<td>it’s all okay.</td>
</tr>
<tr>
<td>S14</td>
<td>it’s so much more fulfilling.</td>
</tr>
<tr>
<td>S15</td>
<td>no, I’m thinking about growing it out actually.</td>
</tr>
<tr>
<td>S16</td>
<td>nurse, I’m really hurting.</td>
</tr>
<tr>
<td>S17</td>
<td>she couldn’t pick up a cup or do anything like that.</td>
</tr>
<tr>
<td>S18</td>
<td>so it’s that word far that’s the difference between the two.</td>
</tr>
<tr>
<td>S19</td>
<td>so let’s put the real line in here.</td>
</tr>
<tr>
<td>S20</td>
<td>so let’s write that down in the form that we’re given in the question.</td>
</tr>
<tr>
<td>S21</td>
<td>so one must count one’s blessings.</td>
</tr>
<tr>
<td>S22</td>
<td>some things people like more and think they’re good changes.</td>
</tr>
<tr>
<td>S23</td>
<td>sometimes it’s just all in, in there or in there.</td>
</tr>
<tr>
<td>S24</td>
<td>that’s small change.</td>
</tr>
<tr>
<td>S25</td>
<td>that’s what I meant.</td>
</tr>
<tr>
<td>S26</td>
<td>that’s, uh, 3 a week, making 6 so far.</td>
</tr>
<tr>
<td>S27</td>
<td>the first go smelled didn’t it?</td>
</tr>
<tr>
<td>S28</td>
<td>they would approach me and say they’re looking for a marriage contract.</td>
</tr>
<tr>
<td>S29</td>
<td>we’ve had 2 to the 6th, 2 to the 5.</td>
</tr>
<tr>
<td>S30</td>
<td>we’ve talked so far about waves on water.</td>
</tr>
</tbody>
</table>

Once the task was set up, it was made available for 30 English-speaking crowdworker students for 26 days in August 2020 and was closed once there had been no responses over a 10-day period. Afterwards, the responses were downloaded and compared to the output of each approach.
To decide whether a sentence was considered “biased” by participants, for each sentence, the following steps were taken:

1. The number of marks for each sentence (i.e., the number of “yes”, “no”, and “unsure”) was computed.

2. The Pearson’s chi-squared test (Pearson, 1900) was applied to each sentence to identify potential statistically significant differences.

3. Sentences received a label based on the highest number of statistically significant marks. For example, if a given sentence received the most “yes” marks and that number was statistically significantly different from the other marks, the sentence was labelled as “biased” based on most marks.

Once the respective label for each sentence was computed, the following metrics were computed: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) for each approach (i.e., Pryzant and Gaucher). Those metrics were chosen for being commonly used to evaluate the performance of classification algorithms by measuring the algorithm’s ability to correctly identify positive and negative instances, as well as the ability to avoid false positive and false negative results (Davis & Goadrich, 2006; Powers, 2020).

4.2.3 RC1: Evaluation and Reflection

This section is divided into four parts. First, it presents the results from the evaluation of the two approaches. Then, key findings are discussed, which are followed by limitations of this research cycle. Finally, core implications for the following study, Study 2 (Chapter 5), are highlighted.

Evaluation of selected approaches

As indicated previously (Section 4.2.2), a sample of 2024 sentences extracted from online videos created by teachers from 91 Open University courses was ultimately
analysed. The Pryzant approach reported that 49.51% (n = 1002) of the analysed 2024 sentences included words that were “subjectively biased”. In contrast, the approach proposed by Gaucher reported that 4.25% (n = 86) of the sentences analysed were inclined towards females and 1.83% (n = 37) towards males, which represents a total of 6.08% (n = 123) of sentences reported as gender-biased (i.e., the sentences contained words considered by Gaucher et al. (2011) as feminine or masculine). Table 4.2 shows examples of those sentences.

Table 4.2: Excerpt of sentences from online learning materials at The Open University identified by Pryzant et al. (2020)’s approach as subjective bias, and by Gaucher et al. (2011)’s as gender bias.

<table>
<thead>
<tr>
<th>Context</th>
<th>Type*</th>
<th>Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arts</td>
<td>S</td>
<td>“it was crazy”</td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>“it’s very hard to understand”</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>(no bias reported)</td>
</tr>
<tr>
<td>Business</td>
<td>S</td>
<td>“Greece disclosed that government debts”</td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>“the way that we bring together different elements”</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>“it was the individual who runs our Gulf of Mexico business”</td>
</tr>
<tr>
<td>Computing</td>
<td>S</td>
<td>“it’s a single complex number”</td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>“it will depend again upon what a customer has in their contract”</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>“how does that help you make a decision about an IT service?”</td>
</tr>
<tr>
<td>Education</td>
<td>S</td>
<td>“all right, boys, we’re going to start off”</td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>“the personality and the behaviour of that child”</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>“to lead, you have to take stock, and step back”</td>
</tr>
<tr>
<td>Health</td>
<td>S</td>
<td>“her hands were literally on her chest”</td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>“the support I believe I need”</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>“this short film is about how judges in court decide sentences”</td>
</tr>
<tr>
<td>Law</td>
<td>S</td>
<td>“we are the true Kurdish people”</td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>“students who engage in these kind of activities”</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>“as a business man how what do you see as the objective?”</td>
</tr>
<tr>
<td>Natural Sciences</td>
<td>S</td>
<td>“identical twins start out as a bizarre accident”</td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>“huge number of child deaths”</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>“because energy, or work, is force times distance”</td>
</tr>
<tr>
<td>Social Sciences</td>
<td>S</td>
<td>“this is perfect”</td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>“the idea of feminism isn’t really something they understand”</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>“one which always involves individual action”</td>
</tr>
</tbody>
</table>

* S = subjective bias; F = female bias; and M = male bias.

To illustrate, Pryzant et al. (2020)’s approach identified sentences such as “we are the true Kurdish people”, and “identical twins start out as a bizarre accident” as biased because the words “true” and “bizarre” are preventing the sentences from being neutral in terms of opinion or feeling. Their algorithm suggested those sen-
tences should be rephrased as “we are the Kurdish people”, and “identical twins start out as an accident” in order to remove the bias.

Similarly, the sentences “it’s very hard to understand”, and “the way that we bring together different elements” are considered by the gender-bias approach female-biased because they contain the words “understand”, and “together”. As highlighted in Section 4.2.2, Gaucher et al. (2011) and previous researchers (also see Section 2.6.2) found evidence suggesting those words are more frequent in female-dominated environments, which suggests that sentences containing those words might be biased towards female individuals. In contrast, the words “decision” and “lead” were reported by them as being more frequent in texts from male-dominated environments, which makes the sentences “how does that help you make a decision about an IT service?”, and “to lead, you have to take stock, and step back” potentially male-biased. Nonetheless, such results should be interpreted with caution as such a corpus of gender-themed words was built in non-educational settings. Section 4.2.3 covers a discussion regarding the limitations of the implemented approaches as well as the implications to future research.

RQ1.1: How do the approaches’ outputs compare to random classification?

Regarding the inferential tests for checking the possibility of random classification, no statistically significant difference of Pryzant’s classifications and random classification was found (it was assumed that a random classifier would classify 50% of sentences as biased and 50% as not biased). In contrast, Gaucher’s inferences were different (statistically significant) from random chance. Those findings are detailed in Table 4.3. To illustrate, the first row indicates that 93.9% of the sentences (n = 1901) were classified by Gaucher’s approach as not biased towards gender (gender bias), and the difference between that proportion and 50% is statistically significant (p-value < 0.05). Indeed, there is no overlapping when comparing the confidence intervals (last columns) of the first rows.

RQ1.2: How do the approaches’ outputs compare to student classifications?
Table 4.3: Possibility of random classification

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Observed Proportion</th>
<th>Test Proportion</th>
<th>P-value (2-tailed)</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender Bias</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>1901</td>
<td>0.939</td>
<td>0.50</td>
<td>0.000</td>
<td>0.928</td>
</tr>
<tr>
<td>Yes</td>
<td>123</td>
<td>0.061</td>
<td>0.051</td>
<td>0.072</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>2024</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subjective Bias</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>1022</td>
<td>0.505</td>
<td>0.50</td>
<td>0.673</td>
<td>0.483</td>
</tr>
<tr>
<td>Yes</td>
<td>1002</td>
<td>0.495</td>
<td>0.473</td>
<td></td>
<td>0.517</td>
</tr>
<tr>
<td>Total</td>
<td>2024</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In terms of human evaluation, 20 participants completed the tasks, and each sentence received at least 10 marks. In particular, 373 marks were returned from the crowd-sourcing platform (AMT). This number represents a response rate of 41% given that 900 assignments were published (30 sentences x 30 expected students). In addition, students took approximately 12 seconds to label each sentence (mean = 11.7; median = 8.0; min = 3.0; max = 53.0; std = 7.9). In terms of bias classification, most sentences were predominantly marked as not biased (n = 29, or 96.7%), while only one sentence (3.3%) was predominantly marked as biased (see Table 4.4). To illustrate, 20% of participants marked the first sentence (i.e., “and I’ll maybe catch up with you again at another time”) as “biased”, where the remainder 80% considered the same sentence as “not biased”. That sentence was also considered by both Pryzant’s and Gaucher’s approaches as “not biased”. In contrast, all participants (Column “No” = 100%) agreed (Column “Agreement” = 1.0) that the sentence “i haven’t put the solar wind in yet, but i will shortly” is not biased (notice that Pryzant’s and Gaucher’s approaches also marked this sentence as “not biased” as indicated in columns 2 and 3 of Table 4.4).

Regarding the implemented approaches, Pryzant’s strategy suggested 15 sentences were biased (50%) and 15 sentences not biased. In contrast, Gaucher’s approach indicated that one sentence (3.3%) was biased, and 29 (96.7%) not biased. Overall, Pryzant’s approach, Gaucher’s approach, and most participants agreed that 15 sentences (50%) were not biased. Regarding the correlation of both approaches when their results were compared to participants’ perspectives, the most accurate approach was Gaucher’s when predicting neutral sentences (93.33%). This means
Table 4.4: Number of marks each sentence received when participants were asked: Is this sentence biased? Columns with % sign indicate the percentage agreement on that item (e.g., 80% of participants indicated that S1 was not biased). The last two columns refer to Pearson’s chi-squared test regarding the differences between proportions.

<table>
<thead>
<tr>
<th>Sentence</th>
<th>#marks</th>
<th>#yes</th>
<th>#no</th>
<th>#unsure</th>
<th>%yes</th>
<th>%no</th>
<th>%unsure</th>
<th>x.squared</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>15</td>
<td>3</td>
<td>12</td>
<td>0</td>
<td>20%</td>
<td>80%</td>
<td>0%</td>
<td>23.40</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>S2</td>
<td>10</td>
<td>1</td>
<td>8</td>
<td>1</td>
<td>10%</td>
<td>80%</td>
<td>10%</td>
<td>14.70</td>
<td>.001</td>
</tr>
<tr>
<td>S3</td>
<td>12</td>
<td>3</td>
<td>6</td>
<td>3</td>
<td>25%</td>
<td>50%</td>
<td>25%</td>
<td>2.25</td>
<td>.325</td>
</tr>
<tr>
<td>S4</td>
<td>12</td>
<td>2</td>
<td>10</td>
<td>0</td>
<td>17%</td>
<td>83%</td>
<td>0%</td>
<td>21.00</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>S5</td>
<td>12</td>
<td>4</td>
<td>7</td>
<td>1</td>
<td>33%</td>
<td>58%</td>
<td>8%</td>
<td>6.75</td>
<td>.034</td>
</tr>
<tr>
<td>S6</td>
<td>13</td>
<td>4</td>
<td>9</td>
<td>0</td>
<td>31%</td>
<td>69%</td>
<td>0%</td>
<td>14.08</td>
<td>.001</td>
</tr>
<tr>
<td>S7</td>
<td>11</td>
<td>0</td>
<td>11</td>
<td>0</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
<td>33.00</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>S8</td>
<td>13</td>
<td>1</td>
<td>12</td>
<td>0</td>
<td>8%</td>
<td>92%</td>
<td>0%</td>
<td>30.69</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>S9</td>
<td>10</td>
<td>3</td>
<td>7</td>
<td>0</td>
<td>30%</td>
<td>70%</td>
<td>0%</td>
<td>11.10</td>
<td>.004</td>
</tr>
<tr>
<td>S10</td>
<td>13</td>
<td>6</td>
<td>7</td>
<td>0</td>
<td>46%</td>
<td>54%</td>
<td>0%</td>
<td>9.92</td>
<td>.007</td>
</tr>
<tr>
<td>S11</td>
<td>12</td>
<td>1</td>
<td>10</td>
<td>1</td>
<td>8%</td>
<td>83%</td>
<td>8%</td>
<td>20.25</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>S12</td>
<td>11</td>
<td>3</td>
<td>7</td>
<td>0</td>
<td>21%</td>
<td>79%</td>
<td>0%</td>
<td>7.64</td>
<td>.222</td>
</tr>
<tr>
<td>S13</td>
<td>14</td>
<td>3</td>
<td>11</td>
<td>0</td>
<td>21%</td>
<td>79%</td>
<td>0%</td>
<td>20.79</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>S14</td>
<td>12</td>
<td>1</td>
<td>10</td>
<td>0</td>
<td>8%</td>
<td>83%</td>
<td>8%</td>
<td>20.25</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>S15</td>
<td>12</td>
<td>3</td>
<td>8</td>
<td>1</td>
<td>25%</td>
<td>67%</td>
<td>8%</td>
<td>9.75</td>
<td>.008</td>
</tr>
<tr>
<td>S16</td>
<td>13</td>
<td>3</td>
<td>10</td>
<td>0</td>
<td>23%</td>
<td>77%</td>
<td>0%</td>
<td>18.23</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>S17</td>
<td>11</td>
<td>6</td>
<td>5</td>
<td>0</td>
<td>55%</td>
<td>46%</td>
<td>0%</td>
<td>8.46</td>
<td>.015</td>
</tr>
<tr>
<td>S18</td>
<td>11</td>
<td>0</td>
<td>9</td>
<td>2</td>
<td>0%</td>
<td>82%</td>
<td>18%</td>
<td>18.27</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>S19</td>
<td>13</td>
<td>3</td>
<td>9</td>
<td>1</td>
<td>23%</td>
<td>69%</td>
<td>8%</td>
<td>12.00</td>
<td>.002</td>
</tr>
<tr>
<td>S20</td>
<td>10</td>
<td>2</td>
<td>7</td>
<td>1</td>
<td>20%</td>
<td>70%</td>
<td>10%</td>
<td>9.30</td>
<td>.010</td>
</tr>
<tr>
<td>S21</td>
<td>11</td>
<td>4</td>
<td>7</td>
<td>0</td>
<td>36%</td>
<td>64%</td>
<td>0%</td>
<td>10.09</td>
<td>.006</td>
</tr>
<tr>
<td>S22</td>
<td>12</td>
<td>1</td>
<td>10</td>
<td>1</td>
<td>8%</td>
<td>83%</td>
<td>8%</td>
<td>20.25</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>S23</td>
<td>13</td>
<td>0</td>
<td>13</td>
<td>0</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
<td>39.00</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>S24</td>
<td>11</td>
<td>2</td>
<td>8</td>
<td>1</td>
<td>18%</td>
<td>73%</td>
<td>9%</td>
<td>11.73</td>
<td>.003</td>
</tr>
<tr>
<td>S25</td>
<td>12</td>
<td>2</td>
<td>9</td>
<td>1</td>
<td>17%</td>
<td>75%</td>
<td>8%</td>
<td>14.25</td>
<td>.001</td>
</tr>
<tr>
<td>S26</td>
<td>11</td>
<td>0</td>
<td>10</td>
<td>1</td>
<td>0%</td>
<td>91%</td>
<td>9%</td>
<td>24.82</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>S27</td>
<td>11</td>
<td>2</td>
<td>9</td>
<td>0</td>
<td>18%</td>
<td>82%</td>
<td>0%</td>
<td>18.27</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>S28</td>
<td>12</td>
<td>3</td>
<td>9</td>
<td>0</td>
<td>25%</td>
<td>75%</td>
<td>0%</td>
<td>15.75</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>S29</td>
<td>13</td>
<td>0</td>
<td>13</td>
<td>0</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
<td>39.00</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>S30</td>
<td>14</td>
<td>1</td>
<td>13</td>
<td>0</td>
<td>7%</td>
<td>93%</td>
<td>0%</td>
<td>33.64</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>

that most participants agreed with Gaucher’s labels regarding the sentences supposed to be neutral (i.e., not biased). In contrast, both approaches were very inaccurate in predicting biased sentences (3.33%). In other words, the sentences considered biased by the two approaches were not supported by the majority of participants, i.e., participants did not agree those sentences were biased. Table 4.6 depicts the accuracy of both methods (based on the label – “yes”, “no”, or “unsure” – with the most marks). To illustrate, [row 2, column 4] indicates that Pryzant’s ap-
Table 4.5: Comparison between approaches and the marks of most participants.

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Gaucher</th>
<th>Pryzant</th>
<th>Participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>S2</td>
<td>no</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>S3</td>
<td>no</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>S4</td>
<td>yes (male)</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>S5</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>S6</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>S7</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>S8</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>S9</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>S10</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>S11</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>S12</td>
<td>no</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>S13</td>
<td>no</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>S14</td>
<td>no</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>S15</td>
<td>no</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>S16</td>
<td>no</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>S17</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>S18</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>S19</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>S20</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>S21</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>S22</td>
<td>no</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>S23</td>
<td>no</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>S24</td>
<td>no</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>S25</td>
<td>no</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>S26</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>S27</td>
<td>no</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>S28</td>
<td>no</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>S29</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>S30</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
</tbody>
</table>

Approach classified 14 sentences (46.67%) as “biased”, but those classifications were not in accordance with the choice of most participants (false positive). In contrast, the second column indicates that both approaches classified one sentence (3.33%) as biased, and that classification was in accordance with the participants’ answers (true positive).
Key findings

The results indicated that both approaches and most participants agreed that half of the sentences were not biased. In contrast, it was also found that there was no major agreement about which sentences were biased. For example, in only three sentences (S7, S23, and S29) the participants who marked them agreed 100% which label it should have, i.e., all participants agreed those sentences were not biased. In contrast, participants had diverging opinions for all other sentences (see Table 4.4). One potential reason for that is the possibility of participants coming from different social groups and, therefore, having different perspectives on what they consider biased. That assumption is in line with previous research regarding the contextual aspect of bias that suggests inter-group bias as context-dependent (see Section 2.4). Accordingly, a lower agreement among participants in a given sentence does not necessarily imply that some participants are wrong regarding such a sentence be biased or not. It may suggest that such sentences are only biased against the social group to which certain participants belong. That also supports the research about inter-group categorisation, which has suggested that people tend to favour members of their own group (in-group) rather than members from other groups (out-group) (Grigoryan et al., 2020; Hewstone & Greenland, 2000; Hogg, 2013; Tajfel et al., 1971). Further details about inter-group categorisation are provided in Chapter 2, Section 2.4.1.

Another aspect of potential bias in online learning materials is the fact that such

Table 4.6: Approaches’ accuracy based on participants’ answers. As there were diverging opinions, “True” means “agreed by most participants”.

<table>
<thead>
<tr>
<th></th>
<th>TP</th>
<th>TN</th>
<th>FP</th>
<th>FN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pryzant et al. (2020)</td>
<td>1</td>
<td>15</td>
<td>14</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(3.33%)</td>
<td>(50.00%)</td>
<td>(46.67%)</td>
<td>(0.00%)</td>
</tr>
<tr>
<td>Gaucher et al. (2011)</td>
<td>1</td>
<td>28</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>(3.33%)</td>
<td>(93.33%)</td>
<td>(0.00%)</td>
<td>(3.33%)</td>
</tr>
</tbody>
</table>

TP: True-Positive; TN: True-Negative; FP: False-Positive; and FN: False-Negative
biases may not necessarily represent a threat to students. For example, it was mentioned in Section 2.5.2 that positive bias in a piece of feedback to students may benefit them (Bader et al., 2019; P. Ferguson, 2011; Harber, 1998; Harber et al., 2012). When considering the approaches implemented in this research cycle, Pryzant’s algorithm suggested that nearly 50% of the sentences analysed are most likely to be subjectively biased (see Table 4.3). While their algorithm does not explain why most sentences are classified as biased, their algorithm was designed to identify bias in terms of language subjectivity, i.e., when the language is skewed in terms of opinion, feeling, or taste.

However, those aspects do not necessarily harm students, for example, the sentence “this is perfect” highlighted in Table 4.2 seems acceptable when the word “this” is referring to “keeping good health” but is unacceptable if referring to “Nazism”. Another example is the sentences “all right, boys, we’re going to start off”, and “we are the true Kurdish people” (see Table 4.2). While they may represent a harmful inclination affecting specific groups (i.e., females and non-Kurdish people may feel excluded from the environment in which the sentences are present), understanding the extent to which those potential biases would harm students seems critical before drawing more precise conclusions.

While broader implications of those findings are discussed in Chapter 7, key implications are presented in the next section. Such implications directly impacted the design of the research cycles that took place during Study 2 of this PhD Project (Chapter 5).

**Limitations**

Despite the results presented in the previous sections, as in any other research, this research cycle is limited in certain aspects. Two main limitations were found, which are related to the data used in this research cycle, i.e., the learning materials, and participant data. More details about each limitation are provided below.
One notable limitation is related to the learning materials used in the research cycle, i.e., the sentences extracted from the OU VLE. While this dataset was vital for answering the RC1 research questions, only a small amount of data was extracted from that platform. Therefore, any generalisation regarding potential biases in that platform (or any other learning environments) should be drawn with caution.

Another important data limitation encompasses the labels provided by participants, i.e., the data collected through the Amazon Mechanical Turk Platform. For instance, this initial research cycle focuses on labels from a general student population. However, the low agreement among them suggests that such recruitment process may benefit from more specific groups of students. In addition, the low response rate (41%) suggests that other recruitment platforms might be worth using.

Despite those limitations, the findings of this research cycle have valuable implications for the next stages of this PhD Project. Such implications are highlighted in the next section.

**Implications for the next research cycle**

In this section, key implications for Research Cycle 2 are shown. These implications are drawn from both the findings and limitations of Research Cycle 1. It is also important to reiterate that the broader implications of this research cycle are presented in the general discussion of this PhD thesis (Chapter 7). The key implications of this research cycle for the following research cycle, RC2, are:

- **Consider the perspectives of particular social groups.** Future studies regarding inter-group bias in text may benefit from focusing on the perspectives of specific groups regarding what they consider ‘biased’. Investigating inter-group bias as a generalised concept might not account for the fact that certain instances of inter-group bias might depend on individual perspectives. In addition, focusing on a general population may only benefit most of that population to the detriment of the perspectives of ethnic minority individu-
als. That assumption was derived from the low agreement among participants when assessing potential biases.

- **Account for contextual aspects when assessing potential biases.** While students’ perspectives seem to affect their judgement regarding what they consider biased, it seems reasonable that the context to which a sentence belongs might also play a critical role when determining what is biased. As highlighted previously, a sentence may have different implications depending on the learning material or course it belongs to.

- **Focus on negative bias.** When uncovering inter-group bias in texts from learning settings, it seems reasonable to understand the actual threat (or benefit) such bias might cause to students. While the approaches implemented in this research cycle seem promising for identifying certain biases in other contexts (e.g., Wikipedia), the findings presented above suggested that some of the potential biases highlighted within the present data are not harmful to students (see Section 4.2.3).

- **Focus on a more diverse corpus of online learning materials.** While the data used in this research cycle led to pertinent insights regarding existing approaches that aimed to identify bias in text, data from more diverse online learning materials might provide additional examples of where potential bias might occur within online learning. That implication is drawn from the limitations regarding the dataset used in RC1 (see Section 4.2.3).

- **Focus on a particular type of inter-group bias.** Future investigations regarding inter-group bias in online learning materials might also benefit from focusing on a particular type of bias (e.g., gender bias, ethnic bias, etc.). For instance, the implications indicated above reinforce the complexity of inter-group bias highlighted in the literature review (see Section 2.4). That assumption is based on the fact that individuals from particular groups might perceive bias differently. That reflection also informed a notable decision of this PhD
project: the focus on ethnic biases in the following research studies.

4.3 Chapter summary

This chapter explored two approaches to uncover whether (or not) it was possible to identify inter-group bias in online learning materials produced by teachers at The Open University. Like many other institutions, teachers at the OU spend considerable time and effort designing inclusive online learning materials. However, with any human-designed activity, there could have been the possibility of unintended biases being introduced at the time those materials were created, when those materials were designed on a large scale. Therefore, two automated approaches drawn from the literature were tested on 2024 sentences sampled from 91 courses across several disciplines. Afterwards, crowd-sourced Higher Education students were asked to complete a task in which they labelled a subset of sentences as biased or not biased. Then, their labels were compared to the automated approaches’ outcomes. While both automated approaches identified potential biases in the analysed sentences, there was no major agreement with the responses from the crowd-sourced students. This suggests that the tested automated mechanisms to identify inter-group bias in non-educational contexts may not be ready for identifying inter-group bias in learning materials from online learning platforms. These insights informed the next research cycles (Chapter 5), which are presented below.
5 | Study 2

5.1 Chapter overview

Chapter 4 highlighted how existing approaches aiming at the identification of group bias in text would perform in online learning materials. For instance, it was found that future approaches for bias identification in online learning materials should focus on the perspective of particular groups, bias beyond word level, negative biases, and bias in large educational data (see Section 4.2.3). This chapter is informed by those challenges and is aimed at understanding how contextual elements affect the identification of ethnic bias in online learning materials and how such biases are perceived by students. In particular, Study 2 was guided by the following underlying research question:

*RQ2.0: How does bias against certain ethnic groups manifest in text-based online learning materials?*

As indicated in Chapter 3 (Methodology), Study 2 comprised two research cycles, which are organised as follows: Section 5.2 (RC2) focuses on designing an online task in which students were asked to assess potential ethnic bias in text-based online learning materials. Section 5.3 (RC3) comprised looking at how White students and students from ethnic minority groups perceive ethnic biases in learning texts. More details about each cycle are provided below.

---

1The research conducted during Study 2 is being prepared for publication. Preliminary details:

- **Albuquerque, J.**, Rienties, B., Holmes, W., & Hlosta, M. (2023). *What is biased, and according to whom? Students' perspectives of ethnic bias in open educational resources*
5.2 Research Cycle 2

Like RC1, RC2 is organised as Analysis and Exploration (Section 5.2.1), Design and Construction (Section 5.2.2), and Evaluation and Reflection (Section 5.2.3). Those stages are detailed below.

5.2.1 RC2: Analysis and Exploration

As indicated in Section 2.6, existing approaches aiming at bias identification had different limitations. For example, they focused on non-educational contexts like bias in Wikipedia (Caverlee, 2022; Dadu et al., 2020; Hube & Fetahu, 2018), they did not consider contextual elements (e.g., subject, sentence, paragraph), and they did not account for the perspectives of particular social groups (e.g., individuals from ethnic minority populations). Accounting for those aspects was critical for the reasons indicated as follows.

First, what is considered biased in non-educational contexts might not be considered biased in educational settings and vice versa. For example, sentences describing an opinion might be considered inappropriate for an encyclopaedia context, where such materials should maintain a high level of neutrality in their articles (Wikimedia Foundation, 2020). In contrast, opinionated content is often considered acceptable and sometimes necessary in educational settings. For example, the sentence “Marriage is a holy union of individuals” (from Table 2.2) could be considered unsuitable for Wikipedia, as the term “holy” imparts a religious connotation to the statement. However, this sentence might be deemed acceptable for religious students who might simply be expressing their beliefs during a class discussion.

Second, existing studies appear to overlook essential contextual factors that may influence the degree of bias in text. For instance, the term “coloured” on its own might be suitable when referring to objects or animals; however, it could be deemed inappropriate when used to describe an individual’s skin colour.
Third, the perspectives of particular groups were not taken into consideration. For example, the study implemented in the previous research cycle (Pryzant et al., 2020) did not specify the population, i.e., the approach focused on what is perceived as biases by Wikipedia’s editors, which can be any individual with access to the Internet. However, the lack of a precise target population makes it hard to understand who considers certain sentences as biased and the extent to which those supposed biases affect particular groups (e.g., individuals from ethnic minority groups). That assumption emerged after reviewing the literature regarding subjective aspects of inter-group bias (Section 2.4.3), which suggested that an individual’s background can affect their judgement of what they consider “biased” against a certain group. Indeed, culture has been shown as a key factor that contributes to shaping the meaning of in-groups and, therefore, the nature of inter-group biases (Brewer & Yuki, 2007; Tajfel et al., 1971).

Accordingly, this research cycle focused on if and how those limitations could be mitigated. In particular, this research cycle looked at the extent to which the design of a bias-identification task could be informed by contextual aspects of online learning materials and, at the same time, how to account for the perspectives of students from different ethnic backgrounds. RC2 was guided by the following underlying question:

*RQ2.1: How do we operationalise contextual and cultural aspects to support the identification of ethnic bias in text-based online learning materials?*

As that research question suggests, this research cycle was centred on methodological aspects of Study 2. In particular, this research cycle focused on defining the source of data (i.e., target population and a platform to extract online learning materials) and designing a labelling task. Defining and evaluating those aspects were important for identifying potential errors before conducting a large-scale intervention. Below, more details about the design of this research cycle are provided.
5.2.2 RC2: Design and Construction

The design comprised defining the source of data (i.e., a place for extracting online learning materials), the target population, and prototyping an online task. The design also included defining an evaluation mechanism in order to inform the decision regarding those aspects (i.e., population, source of data, and task).

Dataset

Online learning materials for this research cycle were extracted from the OER Commons platform\(^2\) – an open educational network where a variety of open educational resources can be created, shared, and modified (“OER Commons & Open Education”, 2007). This platform was chosen for two main reasons: (i) the variety of OERs that could benefit sample representativity; and (ii) the fact that most of its OERs have an open licence, which could make them suitable for downloading, processing, and sharing along with the findings of this study.

Once the data source was established, OERs were selected based on the following selection criteria: (i) Format: only text-based OER were considered; (ii) Language: OERs not in English were excluded; (iii) Licence: only open licence resources were downloaded; (iv) Availability: readily available online, e.g., no authentication needed; and (v) Related to ethnic groups: only sentences containing an explicit reference to an ethnic group as the scope of this research is situated in ethnic differences in online learning. “Explicit reference” refers to a keyword identifying an ethnic group (e.g., “Asian student”, “Black community”, “White woman”). To identify keywords referring to ethnic groups, a non-exhaustive list of terms was built based on the four sets of strings presented in Table 5.1 (S1, S2, S3, and S4).

Accordingly, those keywords were generated based on a synonym strategy in which terms related to Asian, Black, Mixed, and White were included (the next section discusses the choice of those groups). In addition, terms related to “person” and

\(^2\)https://www.oercommons.org/
“location” were also included in the list of keywords. In summary, the final list of keywords comprised the terms in S1 plus the concatenations of terms from the following sets: S2+S3 (e.g., “African academic”, “Afro academic”, “Asian academic”), and S3+S4 (e.g., “academic from Africa”, “academic from Asia”, “academic from China”), and S3 (e.g., “Africans”, “Asians”, “Blacks”).

Table 5.1: Non-exhaustive list of strings (in lowercase) used to create keywords referring to ethnic groups*.

<table>
<thead>
<tr>
<th>S1</th>
<th>africans, asians, blacks, chinese, dark skin, dark-skinned, indians, japoneses, koreans, non-asians, non-blacks, non-whites, white skin, white-skinned, whites</th>
</tr>
</thead>
<tbody>
<tr>
<td>S2</td>
<td>african, afro, asian, black, blackout, brown, caucasian, colored, coloured, darkened, eastern, gray, grey, indian, japanese, korean, non-african, non-asian, non-black, non-white, occidental, oriental, western, white</td>
</tr>
<tr>
<td>S3</td>
<td>academic, acquaintance, administrator, adolescent, adviser, affiliate, alien, allies, ally, american, assistant, audience, aunt, babe, babies, baby, bairn, being, bodies, body, boss, boy, boyfriend, bride, bridgroom, british, brother, buddy, cadet, chap, character, chick, child, chum, citizen, classmate, co-worker, coach, collaborator, colleague, communities, community, companion, comrade, countries, country, creature, cronies, crony, culture, customer, dad, daddies, daddy, damsel, daughter, dependent, director, docent, doctor, dude, educator, electorate, expert, faculties, faculty, families, family, father, fellow, female, fiancé, fiancée, first-year, flatmate, folk, foreigner, freshman, freshmen, friend, gal, gentleman, gentlemen, gentlewomen, girl, girlfriend, graduate, grandfather, grandma, grandmother, grandpa, grannies, group, guest, guy, helpmate, host, household, human, husband, identities, identity, immigrant, individual, infant, inhabitant, inspector, instructor, intellectual, junior, kid, kinsperson, ladies, lady, lassie, leader, learner, lecturer, lecturer, led, local, madam, male, man, manager, matron, member, member, men, mentor, migrant, mom, mommies, mommy, mortal, mother, nation, national, nationalities, nationality, native, neighborhood, neighbour, neighbourhood, nephew, new-born, newborn, niece, officer, official, papa, participant, parties, partner, party, peer, people, person, personalities, personality, personnel, persons, population, pre-schooler, preschooer, president, professional, professor, public, pupil, race, relative, representative, researcher, resident, roomie, roommate, scholar, schoolboy, schoolchild, schoolgirl, schoolmate, schoolteacher, scientist, senior, sibling, sister, societies, society, son, sophomore, soul, spirit, spouse, staff, stepbrother, stepfather, stepmother, stepparent, stepsister, student, supervisor, swain, teacher, team, teammate, technician, teenager, toddler, tot, trainee, trainer, tribe, tutor, twin, uncle, undergrad, undergraduate, visitor, voter, wife, woman, women, worker, youngster, youth</td>
</tr>
<tr>
<td>S4</td>
<td>from africa, from asia, from china, from india, from japan, from north korea, from south korea, from the orient, of africa, of asia, of china, of color, of colour, of india, of japan, of north korea, of south korea, of the orient</td>
</tr>
</tbody>
</table>

Please, note that some terms are considered derogatory.

Initially, 1407 online learning materials were successfully extracted from the OER Commons Platform. Then, after filtering such materials based on the keywords indicated in Table 5.1, 112 online learning materials were selected. Once the mate-
rial were filtered based on those keywords, the next step comprised looking for the specific sentences that also contained those keywords. In essence, after applying the selection criteria indicated above, 345 sentences were ultimately eligible. Finally, from the set of 345 eligible sentences, 20 sentences were randomly selected for this research cycle. That number was defined considering that 20 sentences would not take too long for a labelling task (i.e., participants needed to read all sentences and indicate if they were biased or not). The next section provides more details about the population of this research cycle.

**Population**

The population comprised Higher Education students from the United Kingdom and the United States. Those two countries were chosen because (i) English is the native language for most participants, and (ii) there have been reports of subtle and more explicit differences in how race and ethnicity are discussed in these two contexts (Richardson et al., 2020). In addition, sample selection was based on self-reported ethnicity, in which participants identified themselves as belonging to one of four groups: Asian, Black, Mixed, and White. These groups were defined according to the Ethnicity Harmonised Standard (GSS Harmonisation Team, Accessed 2022, October 14), which aims to ensure consistency and comparability in the collection and reporting of ethnicity data across different sources. Although this categorisation does not perfectly represent all ethnic groups, it was necessary to prevent the sample from being dominated by participants from White-related backgrounds due to census data showing an under-representation of certain ethnic groups. Not blocking the sample based on ethnicity would have resulted in lower sample validity, with a less representative sample of individuals from ethnic minority backgrounds.

A list with the full selection criteria along with the reason for each criterion is shown in Table 5.2.

Regarding the mechanism for recruiting participants, a notable change from RC1 to
Table 5.2: Participants’ selection criteria

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Reason</th>
</tr>
</thead>
<tbody>
<tr>
<td>HE students, i.e., currently enrolled either in an undergraduate degree (BA/BSc/other) or graduate degree (MA/MSc/MPhil/PhD/other).</td>
<td>HE was chosen as the focus of the study due to the prevalence of research on bias in online learning in this context (Section 2.3), providing a strong theoretical foundation.</td>
</tr>
<tr>
<td>Fluency in the English Language.</td>
<td>Participants should be able to judge sentences written in English.</td>
</tr>
<tr>
<td>Belonging to one of the following ethnic groups: Asian, Black, Mixed, or White.</td>
<td>This criterion was added aiming to make the sample more representative as the majority of the UK and US populations come from a White (or similar) ethnic group (based on the UK 2021 Census and the US 2020 Census).</td>
</tr>
<tr>
<td>Adults, i.e., at least 18 years old.</td>
<td>This aimed at facilitating the study implementation as adults are able to make autonomous decisions and provide informed consent to participate in research.</td>
</tr>
</tbody>
</table>

RC2 was the choice of the Prolific Platform \(^3\) instead of AMT. While AMT allowed a quick task setup and had an inexpensively cost for RC1, the particular needs of RC2 made Prolific more suitable for this research cycle. For example, Prolific allows researchers to screen participants before they start the study based on specific criteria (like the one in Table 5.2). To accomplish that, Prolific users are asked to complete a set of demographic questions when they create an account on the platform. Then, when researchers need specific samples (e.g., based on attributes like education level, country, ethnic group, etc.), they only need to specify the sample requirements and the platform automatically sends the study to participants who match those criteria. The pre-screening feature of Prolific benefits research ethics by preventing the “rejection” of participants after they have started the study. In essence, the particular selection criteria shown in Table 5.2 was the main reason for selecting Prolific over AMT.

---

\(^3\)https://prolific.co/
Online task

The online task comprised four main parts: (i) participants’ information sheet (see Appendix C.1); (ii) electronic consent form (see Appendix C.2); (iii) instructions; and (iv) labelling questions. General aspects of the study were included in the information sheet, for instance, the study’s aim, task overview, potential risks and benefits, usage of participants’ data, researchers’ contact, and participant rights to withdraw. After giving their consent, detailed instructions about the task were provided, followed by 23 questions (20 sentences plus 3, attention check, questions, more details about attention check are provided in the next section). Figure 5.1 shows the instructions page and Figure 5.2 shows an excerpt of a labelling question. A general open-ended question was included at the end to collect suggestions or comments from participants.

The task was supported by two core platforms: Prolific and Google Forms. As mentioned previously, Prolific was used to recruit participants based on the selection criteria provided above (see Table 5.2). Google Forms was used for the main task, i.e., the study was listed on Prolific and linked to a questionnaire on Google Forms. The latter was used with the aim of testing potential automation, i.e., Google Forms gives the possibility of generating the questions dynamically by defining a script that loads questions from a spreadsheet. Testing that feature was critical because the next research cycles would contain a larger number of sentences/questions, which would be unpractical to create manually.

Evaluation approach

In order to evaluate this research cycle, several metrics were defined based on the Goal-Question-Metric (GQM) approach (Caldiera & Rombach, 1994), which is a method used to define useful measurements from one or more goals (Dalton, 2019). As its acronym suggests, GQM comprises (i) defining specific goals; (ii) asking questions which will help to achieve the goals; and (ii) defining metrics as a means
Figure 5.1: Task excerpt for RC2 containing the instructions to participants.

of measurement that indicates how to answer the questions. The evaluation approach focused on three main goals: (G1) assess the reliability of data from participants; (G2) assess the suitability of the dataset; and (G3) identify basic requirements for the online task. Once those goals were set, questions and metrics were defined. Table 5.3 summarises the questions derived from those goals and the metrics used to help answer each question. A brief description of each metric is also provided below.

- (M1) Turnout rate. Aimed at identifying whether participants completed the
tasks. This metric was computed based on the number of participants who completed the task (numResponses) and the total number of participants who
Table 5.3: Goals, questions, and metrics that guided RC2.

<table>
<thead>
<tr>
<th>Questions</th>
<th>Metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td>G1: Assess the reliability of data from participants</td>
<td></td>
</tr>
<tr>
<td>Q1: Will participants actually complete the task?</td>
<td>M1: turnout rate</td>
</tr>
<tr>
<td>Q2: Are participants’ answers trustworthy?</td>
<td>M2: attention checks</td>
</tr>
<tr>
<td>Q3: Approval rate in other studies</td>
<td>M3: approval rate in other studies</td>
</tr>
<tr>
<td>G2: Assess the suitability of the dataset</td>
<td></td>
</tr>
<tr>
<td>Q3: Is the dataset suitable for a bias-related study?</td>
<td>M4: Prevalence of bias &gt;20%</td>
</tr>
<tr>
<td>Q4: Would the dataset support a context-based approach?</td>
<td>M5: Number of contextual attributes &gt;0</td>
</tr>
<tr>
<td>G3: Identify basic requirements for the online task</td>
<td></td>
</tr>
<tr>
<td>Q5: How long will the task take?</td>
<td>M6: Avg completion time</td>
</tr>
<tr>
<td>Q6: How many participants per sentence?</td>
<td>M7: # of factors in a factor analysis (PCA)</td>
</tr>
</tbody>
</table>

accepted to take part in the task ($totalParticipants$):

\[
M1 = \frac{numResponses}{totalParticipants} \quad (5.1)
\]

- (M2) Attention checks. Aimed at identifying potentially misleading responses which could diminish the data quality (e.g., if participants respond randomly without paying attention to the question). M2 was computed based on the formula below. Also, Figure 5.3 shows an example of an attention check question.

\[
M2 = \frac{\text{number of attention checks passed}}{\text{total number of attention checks}} \quad (5.2)
\]

- (M3) Approval rate in other studies. This metric was intended to help understand the extent to which participant answers were trustworthy by considering their performance in other studies. This metric is provided by Prolific and indicates the percentage of approval of a given participant when considering all studies in which that participant took place:
Figure 5.3: Example of an attention check question incorporated into the online task.
\[ M_3 = \frac{\text{number of studies approved}}{\text{total number of studies}} \]  (5.3)

- (M4) Prevalence of bias. This aimed to uncover to what extent the dataset would be suitable for this PhD project (i.e., a dataset without potential biases would not be helpful for answering research questions about bias). Therefore, it was initially established that a suitable dataset should have a prevalence of at least 10% bias. That threshold was adopted to guarantee the minimum statistical standards for data analysis in the subsequent research cycles. In other words, 10% of 345 (the size of the OER dataset presented above) is about 34 sentences, which is above the recommended minimum sample size for benefiting the precision and reliability of statistical estimates (Dong & Peng, 2013). Accordingly, M4 was computed based on the number of sentences identified as biased by the majority of participants:

\[ M_4 = \frac{\text{number of sentences considered biased by most participants}}{\text{total number of sentences}} \]  (5.4)

- (M5) Number of contextual attributes. As indicated in the previous research cycle (RC1), a key limitation of existing mechanisms to identify bias in text is their limited context-related attributes, i.e., when investigating bias in text, only sentences were considered without additional information regarding the context from which sentences belong (see Section 2.6.2 for more details). Aiming at addressing such a limitation, the Research Cycle 2 (RC2) sought a dataset from which contextual attributes could be extracted. Contextual attributes are taken to mean additional information that could potentially help individuals assess the extent a sentence was considered biased or not. It is worth reiterating that what is considered biased can depend on context and even culture (see Section 2.4.3). However, the extent to which that assump-
tion would be confirmed for potential inter-group bias in online text-based learning materials is unknown, which led to RQ2.2: How do students from certain ethnic groups perceive ethnic biases in text-based online learning materials? While this research question is covered in the next research cycle (RC3), RC2 concentrated on making sure the chosen dataset contained contextual attributes. Thereby, M5 was considered for assessing the suitability of the dataset. In essence, M5 was defined as:

\[ M_5 = \text{number of contextual attributes} \]  \hfill (5.5)

- (M6) Avg completion time. This aimed at identifying how long participants would take to complete the online task. That metric would allow calibrating the task so that participants would not take too long (or too short) to complete it. Accordingly, M6 comprised the average time participants took to complete the task. In other words, M6 was computed based on the number of participants \((n)\) and the time each participant took to complete the task \((t_i)\):

\[ M_6 = \frac{1}{n} \sum_{i=1}^{n} t_i \]  \hfill (5.6)

- (M7) Number of factors in a principal component analysis (PCA). This metric aimed at estimating how many participants would be necessary to label a given sentence. A PCA approach looks at the minimum number of factors (also known as “components”) that capture most of the variance in the data (Bryman & Cramer, 2011). That assumes that columns with less variance in a dataset have less information. In essence, participant answers were organised in columns and questions in rows, accordingly, the minimum number of columns (participants) that would explain the most variance of the dataset was looked at.
\[ M7 = \text{number of principal components in a factor analysis} \] (5.7)

**Execution**

Once the design was complete, the online task was made available for at least 59 participants on the 9th of November 2021. That number of participants was defined based on a strategy (Viechtbauer et al., 2015) that considers the prevalence of unforeseen problems expected to be captured in a study. Accordingly, at least 59 individuals were necessary to capture unforeseen problems with a prevalence of 5% and 95% confidence.

After the task was made available, the number of participants was constantly monitored. Then, the task was closed when more than 59 responses were recorded, which took approximately 24 hours to receive all responses. Once the task was completed, the data were downloaded for analysis. Afterwards, the data were cleaned, formatted, and the metrics were computed. The results are presented in the next section.

**5.2.3 RC2: Evaluation and Reflection**

A total of 61 participants (self-identified as Asian = 15, Black = 16, Mixed = 15, and White = 15) completed the labelling task. Most participants self-identified as females (females = 50, males = 11), and the majority were currently residing in the UK (UK = 39, US = 22). In addition, most participants were undergraduates (undergrad = 36, grad = 25) at an approximate age of 24-year-old on average (min = 18, median = 22, mean = 24.26, max = 41).

Table 5.5 summarises the metrics. In terms of the metrics related to G1 (assess the reliability of data from participants), the turnout rate was 100%, the approval rate

---

4As indicated in Section 5.2.2 (subsection “Population”), participants were pre-screened using the Prolific platform. As also noted, that process is based on demographic questions completed by participants by the time they sign up for Prolific. Therefore, participants were classified into those groups according to their self-identification.
on other studies was 99.28% (min = 93.33%, median = 100.00%, mean = 99.28%, max = 100.00%), and all participants passed all attention checks.

In terms of the second goal (G2: assess the suitability of the dataset), the estimated amount of bias in the dataset was 40%, which means that 8 out of the 20 sentences were marked as biased by most participants. In addition, the number of contextual attributes initially present in the dataset was 4 (title, type, discipline, and excerpt). A brief description of each attribute is provided in Table 5.4 along with some examples. Those attributes were selected because they were widely available within the dataset, which would allow further data analysis during RC3 (Section 5.3.2 for more details).

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Description</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>The title of the learning material from which the sentence was extracted</td>
<td>“Nelson Mandela &amp; South Africa” “Perspective on the Slave Narrative” “The NGO Handbook”</td>
</tr>
<tr>
<td>Type</td>
<td>The type of the learning material from which the sentence was extracted</td>
<td>Lesson Course Activity</td>
</tr>
<tr>
<td>Discipline</td>
<td>The discipline from which the learning material was extracted</td>
<td>Social Sciences Natural Sciences Mathematics</td>
</tr>
<tr>
<td>Excerpt*</td>
<td>The text excerpt to which the sentence belongs. Usually it comprises the sentence, and its preceding and proceeding sentences.</td>
<td>“More than 350,000 African Americans served in segregated units during World War I, mostly as support troops. Several units saw action alongside French soldiers fighting against the Germans, and 171 African Americans were awarded the French Legion of Honor. [[In response to protests of discrimination and mistreatment from the black community, several hundred African American men received officers’ training in Des Moines, Iowa.]] By October 1917, over six hundred African Americans were commissioned as captains and first and second lieutenants.”</td>
</tr>
</tbody>
</table>

*The corresponding sentence for this excerpt is presented between double brackets (i.e., [[sentence]])

Regarding G3 (identify basic requirements for the online task), the average completion time was 17.63 minutes (min = 4.70; max = 45.77; sd = 9.2). In addition, the
PCA analysis indicated that the variance within participants’ marks was explained by six factors based on a correlation matrix approach (model fit > 94%; the proportion of residuals > 0.05 was less than 43%; residuals were normally distributed based on the Shapiro-Wilk test). Therefore, the PCA approach suggested that at least 6 participants would be necessary to label each sentence.

Table 5.5: Metrics computed in RC2

<table>
<thead>
<tr>
<th>Metric</th>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>Turnout rate</td>
<td>100%</td>
</tr>
<tr>
<td>M2</td>
<td>Attention checks</td>
<td>100% passed</td>
</tr>
<tr>
<td>M3</td>
<td>Approval rate in other studies &gt;90%</td>
<td>99.28</td>
</tr>
<tr>
<td>M4</td>
<td>Prevalence of bias &gt;10%</td>
<td>40%</td>
</tr>
<tr>
<td>M5</td>
<td>Number of contextual attributes &gt;0</td>
<td>4</td>
</tr>
<tr>
<td>M6</td>
<td>Avg completion time</td>
<td>17min 38sec</td>
</tr>
<tr>
<td>M7</td>
<td># of factors in a factor analysis (PCA)</td>
<td>6</td>
</tr>
</tbody>
</table>

Implications for the next research cycles

As indicated in Section 5.2.1, previous studies aiming at identifying bias in text have different limitations. For example, they were conducted in non-educational settings, contextual elements affecting the extent to which a sentence is biased were not accounted, and the perspectives of individuals from particular groups were not considered. Accordingly, this research cycle looked at how those aspects could be operationalised in a bias-identification task. In particular, a target group of participants was initially defined along with a dataset for extracting online learning materials for the task. Then, an online labelling task was designed where the group of participants judged sentences extracted from the online learning materials. To evaluate the quality of participants’ and online learning materials’ data and to test the task setup, a goal-question-metric approach was adopted. The results from this research cycle were used to inform the next ones.

In particular, the results above suggest that recruiting participants through Prolific would be promising for this project. All metrics related to the first goal were pos-
itive, which indicates the data from participants is likely reliable. In addition, the results also suggested the dataset was appropriate for the research scope as the minimum requirements regarding the prevalence of bias and the number of contextual attributes available were met. Regarding the online task, it was found that 6 participants labelling each sentence would be satisfactory for such a task as the variance within labels was explained by 6 participants.

In summary, those findings, along with the artefacts generated in this research cycle (i.e., the dataset and the online task), comprised building blocks for designing the next research cycle (RC3). RC3 focused on understanding the relevance of contextual attributes when uncovering ethnic bias while accounting for the perspectives of students from a variety of ethnic backgrounds. More details about RC3 are provided in the next section.

5.3 Research Cycle 3

While RC2 provided insights about how to design a bias-identification task so that students could assess educational texts for potential ethnic biases, RC3 applied the insights from RC2 to uncover potential ethnic biases in texts from online learning materials and, at the same time, account for the perspectives of students from ethnic minority groups. In particular, RC3 aimed at (i) understanding how individuals from both ethnic minority and White-related ethnic groups perceive ethnic biases; (ii) identifying which contextual elements support the identification of ethnic bias in learning texts; and (iii) uncovering potential causes of ethnic bias in texts by identifying the reasons certain online learning materials are considered biased.

This research cycle is also organised in terms of Analysis and Evaluation (Section 5.3.1), Design and Construction (Section 5.3.2), and Evaluation and Reflection (Section 5.3.3). Those sections are presented below.
5.3.1 RC3: Analysis and Exploration

As indicated in RC1 (Chapter 4), researchers have used machine learning models to automate the identification of ethnic issues based on labelled data extracted from those contexts (Mozafari et al., 2020; Pryzant et al., 2020). While those models have been shown as promising in automating certain classification and predictive tasks, such models are strongly bound to the data used to train them (Nascimento et al., 2018), i.e., they are not tailored to educational settings. Indeed, the results of RC1 suggested that a potential approach to identify inter-group bias in learning settings should consider data from such settings (see 4.2.3). Accordingly, the previous research cycle (RC2) looked at a potential source of data that would include contextual aspects of online learning materials.

Once a suitable source of data was identified, this research cycle (RC3) focused on labelling the full dataset identified in RC2 so that the aims indicated in the previous section could be achieved. In addition, a larger labelled dataset would be needed to design the approach to identify ethnic bias in educational settings that is described in the subsequent study (i.e., Study 3, Chapter 6). In essence, RC3 focused on answering the following research questions:

- RQ2.2: How do students from certain ethnic groups perceive ethnic biases in text-based online learning materials?
- RQ2.3: To what extent do contextual attributes extracted from text-based online learning materials support the identification of potential ethnic bias?
- RQ2.4: What makes text-based online learning materials potentially biased or not biased according to students?

Overall, the answer to each research question comprised scaling up the previous intervention to generate a more robust dataset to be used in Study 3 (i.e., when different computational models were tested to understand how they would perform in a bias-identification task). Accordingly, this research cycle is informed not only
by the findings of RC1 (limitations of existing bias-detection approaches) but also by the evidence produced in RC2 (operationalisation of a bias-labelling task). More details about the methodological approach used to answer those research questions are provided in the next section (Design and Construction).

5.3.2 RC3: Design and Construction

As indicated above, the design of this research cycle incorporated elements from the two previous interventions. For example, this research cycle also comprised an online task where participants were asked to label a dataset. Below, details about the sample and online task are presented.

Data and population

In contrast to the previous research cycle, RC3 comprised a larger intervention where 325 sentences (the remainder of the dataset) were labelled by 132 participants (66 from ethnic minority backgrounds, and 66 from White-related backgrounds). As completing a labelling task with 325 sentences would require substantial effort from participants, those sentences were randomly split into 22 batches (i.e., 15 sentences per batch). Then, each batch was labelled by 6 participants (3 from each group). That number of participants per batch was defined based on the findings of RC2, which suggested that each sentence should be labelled by at least 6 students (see Section 5.2.3). Below, more details are provided regarding the online task and the data analysis.

Online task

Like RC2, the online task comprised the following sections: (i) participants’ information sheet (see Appendix C.1); (ii) electronic consent form (see Appendix C.2); (iii) instructions (see Appendix C.3); and (iv) labelling questions (see Appendix C.4). In addition, for each labelling question, an open-ended text box was added aiming
at collecting the reasons why the respective sentence was considered biased (or not biased). An example of that open-ended item is also provided in Appendix C.4.

Furthermore, as this research cycle also focused on contextual attributes, a fifth section containing context-related questions was also added. Those context-related questions comprised 5-point Likert-style items where participants were oriented to rate the importance of the four OER attributes presented with each sentence (title, type, discipline, and excerpt). For instance, they were instructed: “Based on the previous questions, please indicate the importance of the following elements when deciding if the sentences were biased or not” (see Appendix C.4.2).

Data analysis pipeline

The data analysis comprised two main parts: quantitative and qualitative. The quantitative analysis covered the research questions “RQ2.2: How do students from certain ethnic groups perceive ethnic biases in text-based online learning materials?” and “RQ2.3: To what extent do contextual attributes extracted from text-based online learning materials support the identification of potential ethnic bias?” In contrast, the qualitative analysis concentrated on answering “RQ2.4: What makes text-based online learning materials potentially biased or not biased according to students?” Below, the steps taken in each analysis are detailed.

Quantitative. To answer “RQ2.2: How do students from certain ethnic groups perceive ethnic biases in text-based online learning materials?” the number of sentences marked by the majority of students as “biased”, “not biased”, and “unsure” were computed. Then, the marks in terms of the ethnic group were summarised, i.e., the number of marks each label received by individuals from ethnic minority groups (Asian, Black, and Mixed) and by individuals from a White-related background. In addition, examples of sentences indicated as “biased” by each group were also provided. Regarding “RQ2.3: To what extent do contextual attributes extracted from text-based online learning materials support the identification of po-
tential ethnic bias?” the importance that participants gave to contextual attributes (i.e., title, type, discipline, and excerpt) was looked at. That was achieved by summarising the scores each attribute received and conducting an analysis of variance (ANOVA).

**Qualitative.** As indicated above, this analysis aimed to answer the research question “RQ2.4: What makes text-based online learning materials potentially biased or not biased according to students?” Accordingly, it looked for patterns in the comments provided by participants in the open-ended question, i.e., the question they were asked to comment on why they think a given sentence is (not) biased (see Appendix C.4). As indicated in Section 3.4.1, Thematic Analysis was the chosen approach for capturing those patterns given its ability to enable researchers to gain insights into the perspectives of participants in a systematic way (Flick, 2022). That systematic process was based on the step-by-step guide for conducting thematic analyses provided by Braun et al. (2017), which comprises the following steps:

1. **Familiarisation with the data:** an immersive engagement with the data to develop a deep understanding of the content.

2. **Generating initial codes:** identifying relevant elements in the data and assigning preliminary codes.

3. **Searching for themes:** aggregating codes to uncover potential themes.

4. **Reviewing themes:** a critical evaluation of the identified themes to ensure they are supported by the data.

5. **Defining and naming themes:** Refining and clearly defining the themes, assigning descriptive labels.

6. **Producing the report:** Generating a report that presents the themes and supports them with data excerpts.

Given the substantial volume of participant comments (n = 1158), a pragmatic ap-
proach was adopted to manage the scope of the analysis so that it could be feasible for this PhD project. Accordingly, a stratified random sample of 100 comments was extracted for detailed analysis in which 50 comments were selected from items marked as “biased” and 50 comments from items marked as “not biased”. This split aimed at a balanced representation of perspectives while accounting for comments on bias and non-bias.

5.3.3 RC3: Evaluation and Reflection

The evaluation is organised according to each research question.

RQ2.2: How do students from certain ethnic groups perceive ethnic biases in text-based online learning materials?

The findings unveiled that 75 (21.7%) sentences were marked by most participants as “biased”, 234 (67.8%) as “not biased”, and 36 (10.4%) as “unsure”. This means that, although most of the analysed sentences were not considered biased, more than 1 in 5 sentences were considered inappropriate by most participants. In addition, when checking most marks within ethnic groups, it was found that 77 sentences (22.3%) were considered biased by students from an ethnic minority group, and 88 (25.5%) were flagged “biased” by students from other groups (i.e., White-related ethnic groups).

The results also suggested a statistically significant disparity of sentences exclusively denoted as biased by each group. For instance, 39 (11.3%) sentences were exclusively considered biased by ethnic minorities, which is statistically significant when contrasted with the remaining 88 sentences (25.5%) marked as biased ($X = 22.2$; $p$-value < .001). In other words, approximately 1 in 10 sentences were considered inappropriate by students identified as Asian, Black, or Mixed, but considered appropriate by students identified as White.

In contrast, 50 (14.5%) sentences were exclusively marked as biased by White stu-
Figure 5.4: Sentences designated “biased” by Group A (individuals from ethnic minority groups) and Group B (White individuals).

(* Sentences deemed “biased” by the majority (> 50%) of Group A or Group B.)

Students, which is statistically different from the remainder of 77 (22.3%) sentences (X = 6.5; p-value = .01). In other words, most students from ethnic minorities backgrounds did not agree with the 14.5% of sentences identified as biased by most White students. This and the previous results are illustrated in the diagram in Figure 5.4.

Furthermore, Table 5.6 provides some examples of sentences indicated as biased by most participants from certain groups. For example, the sentence “These documents chronicle a case in the wider wave of violence that targeted people of colour during Reconstruction” was considered biased by most students from ethnic minority groups, but it was not considered biased by individuals from over-represented ethnic groups. Another example is the sentence “Tobias Gibson lamenting how out of control blacks on his plantation has become, 1864”, which was considered biased by most participants.
Table 5.6: Examples of sentences considered biased by most individuals from an ethnic minority background (A), White participants (B), and by the majority of both groups (C).

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Keyword</th>
<th>Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>“It was the white man’s burden to lift up their lesser brethren through benevolent colonisation”</td>
<td>white man</td>
<td>A</td>
</tr>
<tr>
<td>“It took the vote away from everyone but white males worth 50 pounds”</td>
<td>white males</td>
<td>A</td>
</tr>
<tr>
<td>“Separate schools were organized for black children”</td>
<td>black children</td>
<td>A</td>
</tr>
<tr>
<td>“These documents chronicle a case in the wider wave of violence that targeted people of color during Reconstruction”</td>
<td>people of color</td>
<td>A</td>
</tr>
<tr>
<td>“Levittown contracts contained covenants restricting occupancy by members other than the Caucasian race”</td>
<td>Caucasian race</td>
<td>A</td>
</tr>
<tr>
<td>“More than one fifth of the South’s adult white male population (some 260000) was lost fighting for the Confederacy”</td>
<td>white male</td>
<td>B</td>
</tr>
<tr>
<td>“Describe the role free blacks played in Texas during reconstruction”</td>
<td>blacks</td>
<td>B</td>
</tr>
<tr>
<td>“Another consideration in post-bellum America was a new question to southern society: What would be the role of the newly freed black population of the South?”</td>
<td>black population</td>
<td>B</td>
</tr>
<tr>
<td>“Both blacks and whites are in this long line”</td>
<td>blacks</td>
<td>B</td>
</tr>
<tr>
<td>“In black families is the mother always dominant?”</td>
<td>black families</td>
<td>B</td>
</tr>
<tr>
<td>“Tobias Gibson lamenting how out of control blacks on his plantation have become, 1864”</td>
<td>blacks</td>
<td>C</td>
</tr>
<tr>
<td>“The fact the college president invites black people to his party”</td>
<td>black people</td>
<td>C</td>
</tr>
<tr>
<td>“What were two important provisions white Democrats included in the 1876 constitution that disenfranchised most blacks?”</td>
<td>blacks</td>
<td>C</td>
</tr>
<tr>
<td>“The fact they have black women computers hired to this important work”</td>
<td>black women</td>
<td>C</td>
</tr>
<tr>
<td>“Freeborn blacks could stay in Virginia but emancipated African-Americans were generally required to leave the state”</td>
<td>blacks</td>
<td>C</td>
</tr>
</tbody>
</table>

RQ2.3: To what extent do contextual attributes extracted from text-based online learning materials support the identification of potential ethnic bias?

In general, most OER contextual attributes were considered important by participants (mean >3 on a scale from 1 to 5). In particular, the text surrounding each sentence (excerpt) was considered the most important when assessing ethnic bias (mean = 4.22, sd = 1.04), followed by discipline (mean = 3.36, sd = 1.33), type (mean = 3.17, sd = 1.41), and title (mean = 2.82, sd = 1.18). The one-way ANOVA performed to compare the importance of such attributes when assessing ethnic bias revealed a statistically significant difference between most of them ($F(3, 133) = 30.947, p < .001$). The Tukey’s test for multiple comparisons indicated the per-
ceived importance was statistically different between Excerpt-Discipline (diff = .854; 95% CI[.465, 1.243]; p<.001), Title-Discipline (diff = -.540; 95% CI[-.929, .151]; p =.002), Title-Excerpt (diff = -1.394; 95% CI[-1.783, 1.005]; p<.001), and Type-Excerpt (diff = -1.051; 95% CI[-1.440, .662]; p<.001). The same test suggested no statistically significant differences between Type-Discipline (diff = -.197; 95% CI[-.586, .192]; p = .559), and Type-Title (diff = .343; 95% CI[-.046, .732]; p=.105).

RQ2.4: What makes text-based online learning materials potentially biased or not biased according to students?

As indicated in Section 5.3.2, to answer this research question, a thematic analysis of participants’ comments was conducted (i.e., the comments in which participants indicated why a given sentence was biased or not). Accordingly, the findings revealed that they gave many reasons for sentences being biased. As also indicated in Section 5.3.2, those reasons were organised into themes and are summarised in Table 5.7.

Some themes were more frequent than others. For instance, the most frequent theme was poor terminology, i.e., the application of keywords or expressions to represent an ethnic group or individual. For example, the sentences “Describe the lives of free blacks and the laws that limited their freedom and economic opportunities” and “Do the illustrations depict non-whites in subservient and passive roles or in leadership and action roles”.

Sentences’ tone or connotation leading to a false premise (misleading statement) was another explanation participants gave to justify their assumption of bias. For example, when analysing the sentence “The fact the college president invites black people to his party”, a participant commented “Why should black people being invited be something shocking?”.

In addition, other students remarked that some sentences showed an explicit prefer-
Table 5.7: Description of themes that emerged from the Thematic Analysis

<table>
<thead>
<tr>
<th>Bias?</th>
<th>Theme/Reason</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>Poor Terminology</td>
<td>The terminology used to represent an ethnic group or individual was inappropriate.</td>
</tr>
<tr>
<td>Yes</td>
<td>Misleading Statement</td>
<td>The sentence leads to a false premise (e.g., when it contains false information, when the tone is considered inappropriate)</td>
</tr>
<tr>
<td>Yes</td>
<td>Ethnic Preference or Segregation</td>
<td>Explicit preference or segregation towards/against a particular ethnic group or individual.</td>
</tr>
<tr>
<td>Yes</td>
<td>Generalisation</td>
<td>The sentence is generalising a negative fact to an entire ethnic group.</td>
</tr>
<tr>
<td>Yes</td>
<td>Unnecessary Specification</td>
<td>A particular aspect of an ethnic group or individual that is not relevant to the discussion is mentioned.</td>
</tr>
<tr>
<td>Yes</td>
<td>No Reason</td>
<td>No particular reason was given.</td>
</tr>
<tr>
<td>Yes</td>
<td>Negative Language</td>
<td>The sentence contains negative or diminishing language.</td>
</tr>
<tr>
<td>No</td>
<td>Neutral Or Not Offensive Language</td>
<td>The sentence is neutral or not offensive.</td>
</tr>
<tr>
<td>No</td>
<td>Factual Statement</td>
<td>The sentences is describing a fact (e.g., a historical event)</td>
</tr>
<tr>
<td>No</td>
<td>Context-Exemption</td>
<td>The sentence is not considered biased because of its context.</td>
</tr>
<tr>
<td>No</td>
<td>No Reason</td>
<td>No particular reason was given.</td>
</tr>
<tr>
<td>No</td>
<td>Group-Exemption</td>
<td>The target ethnic group or individual is not subject to ethnic bias because their ethnic group has no history of discrimination.</td>
</tr>
</tbody>
</table>

ence or segregation towards/against particular groups. For example, one participant marked the sentence “The lines between refined white womanhood and degraded enslaved Black femaleness were no longer so clearly defined” as biased and commented “The sentence implies white women cannot be underprivileged in any way and ‘white’ is not capitalised like ‘Black’.”

An equal number of mentions suggested improper generalisations, e.g., generalising a negative fact to an entire group. For example, the sentences “Many free blacks were able to become businessmen and leaders” and “Therefore not only were more white males allowed to vote but that vote also had a direct effect on the outcome of presidential elections”.

Other comments suggested unnecessary specifications, e.g., sentences citing individuals’ skin colour where it was irrelevant to the context. For example, the sentence “About twelve years ago I hired a whaleboat and four black men and proceeded to LongIsland after a load of round clams” in which a participant commented “I don’t
know why black men is mentioned specifically when the excerpt refers to them later as just men”.

Some participants noted the use of negative language. For example, the sentence “The franchise or right to vote was being extended to more white males as income-related eligibility requirements were being dropped by more states” in which a participant commented “I feel like it is negatively biased on the grounds that they mention income. It feels as though they are negative about the fact that not every 'white male' would make a lot of money.”.

Finally, other participants did not specify why they indicated sentences as biased. For example, a participant commented “Although the sentence is bias[ed] against white people the contents of it are definitely true and needed to be said” after marking a sentence as biased, for instance, the sentence “He believed that the United States governments both state and federal represented the interests of whites exclusively which made political participation by black Americans a waste of time at best and complicity in their own oppression at worst”.

Figure 5.5 summarises those themes along with some comments from participants.
Figure 5.5: Patterns (or “themes”) that emerged from participant comments after the Thematic Analysis.
In terms of reasons why participants judged sentences as not biased, the results suggested five main categories: neutral or not offensive language, factual statement, context-exemption, group-exemption, and no reason. **Most comments suggested the language used was not biased because it was not negative or offensive**, e.g., they said the sentences presented fair descriptions of ethnic groups or contexts. Other participants commented that some sentences were merely describing or explaining a fact without causing any harm to anyone. Similar claims were found on comments classified as “context-exemption”, where participants argued sentences were not biased because of the historical context or the material type in which they were found, e.g., passages from old textbooks. A smaller number of comments suggested there was no bias because the target group or individual had not suffered ethnic discrimination in the past, e.g., sentences referring to individuals belonging to a White-related ethnic group. There were also comments where participants did not provide a specific reason.

While the implications of those findings for research beyond this PhD project are discussed in Chapter 7, key implications informing the next research cycle are discussed below.

**Implications for the next research cycle**

In this research cycle, how ethnic bias is perceived by students from different ethnic groups in online text-based learning materials extracted from online platforms was looked at. Higher Education students from different ethnic groups assessed each sentence for potential ethnic bias. The results suggested there was bias in more than 20% of the dataset. However, the results also revealed no consensus on what students perceived as biased, i.e., only 30% of all sentences were considered biased by individuals of both groups (i.e., students from ethnic minority groups and White students). When participants were asked about the importance of contextual elements that could help them identify ethnic bias in a given sentence, they expressed
the sentences surrounding text (i.e., the excerpt to which a sentence belongs, see Figure 5.2) as the most valuable attribute. Furthermore, sentences were considered biased for different reasons, e.g., “poor terminology” was the most common indication of ethnic bias. In contrast, “adequate language” was the primary reason for sentences remarked as “not biased”. Below, those key findings are discussed as well as their implications for the next research cycle.

One key finding of this research cycle is that, although the total amount of ethnic bias perceived by each group was about the same (22.3% and 25.5%), a statistically significant number of sentences was perceived as not being biased by both populations (i.e., 39 and 50 in Figure 5.4). In other words, most individuals from each group did not agree with all sentences considered biased by the other group. This finding suggests the existence of online learning materials that are considered biased by students identified as Asian, Black, or Mixed, but not considered biased by students identified as White (and vice versa). That confirms the need to adopt data that account for the perspective of different ethnic groups when conducting the next research cycle.

Results also highlighted the importance of contextual elements (i.e., title, type, discipline, and excerpt) when dealing with ethnic biases in online learning materials. For instance, students indicated the surrounding text of sentences (the “excerpt”) as the most important element when assessing bias. This finding has direct implications for potential approaches aiming to identify ethnic biases in online learning materials (both human-based and computer-based). For example, such attributes could be used to estimate the sentence’s context when creating machine-learning models to predict ethnic bias.

The results also revealed that students considered sentences either biased or not biased for different reasons. For example, most of the analysed comments suggested sentences were biased because of the terminology used, i.e., participants considered that the language used to refer to an ethnic group was not appropriate. Accordingly,
incorporating such terminology in predictive models seems promising when looking for a bias-identification approach. Similarly, students gave different reasons for the instances they considered not biased. For instance, most comments suggested sentences were not biased because they were not offensive or not negative. Those findings also suggest the incorporation of such elements when creating new models (e.g., the valence of a sentence, which indicates the extent to which the sentence is positive or negative) can be helpful.

Overall, these findings were used to inform the next research cycle, which will be presented in the next chapter.

5.4 Chapter summary

By asking students from both ethnic minority and White-related ethnic backgrounds to label a dataset with more than 300 contextualised sentences, the two research cycles presented in Study 2 provide an initial baseline of how ethnic biases manifest in online learning materials. Overall, the results confirmed a statistically significant difference in what is perceived as “ethnic bias” by students from ethnic minority and White backgrounds although there were major differences. Furthermore, the results suggest the use of contextual elements is fundamental when assessing ethnic bias in text, in particular, the excerpt to which a sentence belongs. In line with these findings, the next research cycles directly benefit from this study by using the dataset as a baseline to test different models aiming at identifying ethnic issues in learning settings.
6 | Study 3

6.1 Chapter overview

In the previous chapter (Chapter 5), a baseline for how ethnic biases look in online learning material was built based on student perspectives. That baseline (i.e., a labelled dataset) could potentially enable the extraction of meaningful features to be used by supervised machine learning methods. By doing this, it could help understand to what extent these supervised methods could assist in the identification of bias in educational texts. Accordingly, in Study 3, commonly used supervised methods were applied to the labelled dataset from Study 2, and their performance was checked against the identification of perceived ethnic bias in text-based OERs. In essence, this study focused on answering the underlying research question:

*RQ3.0: To what extent can Learning Analytics assist in the identification of ethnic bias in text-based online learning materials?*

In particular, Study 3 was divided into two research cycles: RC4 and RC5. RC4 aimed to select useful features for input into classification algorithms. The choice of those features was guided by previous studies that had investigated aspects of language leading to potential biases, for example, the extent to which a text is negative (or positive), clues of the author’s emotional state, complex sentences leading to subjective language, among others (see Section 6.2.2 for details). Once potential features were identified, RC5 focused on testing commonly used classification methods against the task of ethnic bias identification in learning texts. Accordingly, RC5 used both the dataset from RC3 (see Section 5.3.3) and the potential features from RC4. More details about each research cycle are provided below, beginning

---

1The research conducted during Study 3 is being prepared for publication. Preliminary details:

with RC4.

### 6.2 Research Cycle 4

#### 6.2.1 RC4: Analysis and Exploration

RC4’s analysis comprised exploring potential features commonly used in text classification that could be extracted from the dataset built during the previous research cycle (RC3). In particular, RC4 focused on answering the following research question:

*RQ3.1: Which features might support the identification of ethnic bias in text-based online learning materials?*

To answer that question, a feature selection approach was conducted. Below, features that could potentially assist the task of identifying ethnic bias in texts are explored.

**Potential features for detecting ethnic bias in text**

Different aspects of language have already been used to analyse and explore text content. This research cycle explores some aspects that could be relevant for identifying ethnic bias in text-based online learning materials. With the aim of improving the logical flow of this research cycle, the aspects are organised into five categories: psycholinguistic markers, linguistic abstraction, language valence, ethnic group mentions, and contextual attributes. Their definition and rationale for why they seemed promising as potential features are provided below.

**Psycholinguistic markers** (*psychological*) comprise “*indicators that reflect the psychological state of the author at the time of writing the text*” (Sboev et al., 2015, p. 308). Furthermore, Sboev et al. (2015) argued that those markers can allow the determination of emotive texts that reflect the author’s level of excitement. Accord-
ingly, psycholinguistic markers were chosen as potential features based on their potential relation to the author’s emotional state, which plays an essential role in motivating bias (see Section 2.4.1). In essence, the works of Tajfel (e.g., Tajfel (1970) and Tajfel et al. (1971, 1979)) suggested that inter-group bias is related to the way individuals place themselves into social groups that have emotional significance and value to them. Therefore, the following psycholinguistic markers were used:

- Aggressiveness (aggressiv): a marker that can suggest when the author is “excited and ready for immediate action” (Sboev et al., 2015, p. 309) and is computed based on the number of verbs and words in a text.

- Socialisation (socialisation): this coefficient indicates readiness to action, which can demonstrate the level of human socialisation (Sboev et al., 2015). It is computed based on the number of verbs and nouns in a text.

- Stability (emo_stability): according to Sboev et al. (2015), this suggests the author’s level of emotional stability (e.g., it can indicate emotional unrest or anxiety). This marker is computed based on the number of verbs and adjectives in a text.

- Self-reference ratio (self_ref): this marker comprises the number of first-person pronouns divided by the total number of pronouns in a text. It was included based on the assumption that when authors use self-reference, they can favour themselves (or their own group), which is suggested by the in-group/out-group dynamics presented in Section 2.4.

Detailed computations for each psycholinguistic marker can be found in Table 6.1.

Another aspect of language that could suggest potential biases in text is linguistic abstraction (abstraction), which “can be used to characterise how people select verbs and adjectives to describe a person or a behavioural event at different levels of description, ranging from concrete to abstract” (Tincher et al., 2016, p. 349). In
Table 6.1: List of psycholinguistic features and how they were computed.

<table>
<thead>
<tr>
<th>Category</th>
<th>Feature</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>psychological</td>
<td>Aggressiveness</td>
<td>$aggressiv = \frac{\text{number of verbs}}{\text{number of words}}$</td>
</tr>
<tr>
<td>psychological</td>
<td>Emotional Stability</td>
<td>$emo_{\text{stability}} = \frac{\text{number of adjectives} + \text{number of adverbs}}{\text{number of nouns} + \text{number of verbs}}$</td>
</tr>
<tr>
<td>psychological</td>
<td>Self-Reference Ratio</td>
<td>$self_{\text{ref}} = \frac{\text{number of 1st person pronouns}}{\text{total pronouns}}$</td>
</tr>
<tr>
<td>psychological</td>
<td>Socialisation</td>
<td>$socialisation = \frac{\text{number of verbs}}{\text{number of nouns}}$</td>
</tr>
</tbody>
</table>

Figure 6.1: Image from Tincher et al. (2016) based on the work of Mass and colleagues illustrating how language abstraction can be used to describe the behaviour of person A. In this example, the number next to each sentence indicates how abstract the sentence is (i.e., 1 means concrete language, and 4 means abstract language).

Accordingly, Maass (1999) found that in-group members tend to use abstract terms (e.g., adjectives) to describe the positive behaviours of members of their own group and concrete terms (e.g., action verbs) to describe their negative behaviours. Similarly, they also identified that when describing the behaviours of out-group members, concrete terms are used to describe positive behaviours and abstract terms to describe negative behaviours.
describe negative behaviours. These are summarised in Table 6.2.

Table 6.2: How the choice of abstract and concrete terms (e.g., adjectives, verbs) can make language biased when describing the behaviours of in-group members and out-group members.

<table>
<thead>
<tr>
<th>Behaviour</th>
<th>In-group</th>
<th>Out-group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative</td>
<td>Concrete language</td>
<td>“A is hitting the other person”</td>
</tr>
<tr>
<td>Positive</td>
<td>Abstract language</td>
<td>“A is considerate”</td>
</tr>
</tbody>
</table>

In addition to the use of verbs and adjectives as potential indicators of abstraction, researchers have also shown that the variety of words and parts of speech – also known as lexical diversity (Gregori-Signes & Clavel-Arroita, 2015) – can also indicate subjective and abstract language (Wiebe et al., 2004). Given the potential relation of language abstraction to biases in text, the following features were also selected (see Table 6.3 for details):

- Adjective ratio (adj_ratio): computed based on the total number of adjectives in a text, which could give some clues about abstract language (Maass, 1999).
- Verb ratio (vrb_ratio): calculated according to the number of verbs in a text and could indicate concrete language (Maass, 1999).
- Sentence length (sent_len): total number of words in a sentence, which suggests how complex (i.e., abstract) a text is.
- Unique word ratio (unique_word): similar to the previous one, but only accounts for unique words.
- Unique part of speech ratio (unique_pos): the number of unique parts of speech (PoS) divided by the total number of PoS.
- Adverb ratio (adv_ratio): based on the number of adverbs in a text.
- Noun ratio (noun_ratio): based on the number of nouns in a text.
- Pronoun ratio (pnoun_ratio): based on the number of pronouns in a text.
Table 6.3: Features related to linguistic abstraction and their respective formulas.

<table>
<thead>
<tr>
<th>Category</th>
<th>Feature</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>abstraction</td>
<td>Adjective Ratio</td>
<td>( \text{adj_ratio} = \frac{\text{Number of adjectives}}{\text{Number of PoS}} )</td>
</tr>
<tr>
<td>abstraction</td>
<td>Adverb Ratio</td>
<td>( \text{adv_ratio} = \frac{\text{Number of adverbs}}{\text{Number of PoS}} )</td>
</tr>
<tr>
<td>abstraction</td>
<td>Noun Ratio</td>
<td>( \text{noun_ratio} = \frac{\text{Number of nouns}}{\text{Number of PoS}} )</td>
</tr>
<tr>
<td>abstraction</td>
<td>Pronoun Ratio</td>
<td>( \text{pnoun_ratio} = \frac{\text{Number of pronouns}}{\text{Number of PoS}} )</td>
</tr>
<tr>
<td>abstraction</td>
<td>Sentence Length</td>
<td>( \text{sent_len} = \frac{\text{Number of words}}{\text{Number of PoS}} )</td>
</tr>
<tr>
<td>abstraction</td>
<td>Unique PoS ratio</td>
<td>( \text{unique_pos} = \frac{\text{Number of unique PoS}}{\text{Number of PoS}} )</td>
</tr>
<tr>
<td>abstraction</td>
<td>Unique Word Ratio</td>
<td>( \text{unique_word} = \frac{\text{Number of unique words}}{\text{sent_len}} )</td>
</tr>
<tr>
<td>abstraction</td>
<td>Verb Ratio</td>
<td>( \text{verb_ratio} = \frac{\text{Total verbs}}{\text{Number of PoS}} )</td>
</tr>
</tbody>
</table>

Another aspect considered in this research cycle as a promising feature for identifying bias in educational texts was **language valence** (valence). Valence depicts one dimension of emotion and refers to the extent to which a stimulus is positive or negative (Osgood et al., 1957). In language, valence can be realised in words and phrases as a result of social exclusions (Mendelsohn et al., 2020). For example, individuals from certain target groups (e.g., ethnic minority populations) are sometimes associated with negative characteristics to exclude them from “the realm of acceptable norms and values” (Mendelsohn et al., 2020, p. 4). Therefore, as ethnic bias can be related to group discrimination (see Section 2.4), text valence (valence) was hypothesised as a potential feature for a bias-identification task.

As shown in Chapter 2 (Section 2.4), group categorisation is a primary element of inter-group bias. Based on that assumption, it was hypothesised that **ethnic group mentions** (mentions) in text could be related to a sentence being biased against (or in favour of) that group (see Table 5.1 for the list of keywords used). Accordingly, the following features were also considered relevant for this research cycle:

- Mentions to ethnic minority groups (ethnic_min), which comprised the number of keywords in a sentence referring to ethnic minority populations.
- Mention to White-related populations (ethnic_maj): Similar to the above, this comprised keywords referring to numerically dominant populations (i.e., White-related ethnic groups).
• Target group (target_ethnic): this comprised the keyword referring to the target group being analysed in the sentence. As indicated at the beginning of this chapter (Section 6.1), the dataset used in this research cycle was labelled during RC3. Accordingly, when judging each sentence, participants were asked to consider a target group (i.e., they were asked to judge whether a sentence was biased against (or in favour of) a certain ethnic group being explicitly referred to). Section 5.3.2 provided more details about the labelling task.

Finally, contextual attributes (context) comprised characteristics of online learning materials that could indicate the context from which they were extracted. Those attributes were the type of the online learning material (e.g., lesson plan, activity), the title of the online learning material, the discipline from which a given material was extracted (e.g., social sciences, mathematics), and the excerpt to which a given sentence belongs. Those attributes were chosen given the subjective nature of ethnic bias in which what is considered biased depends on contextual aspects (see Section 2.4). Therefore, it was hypothesised that contextual attributes would be relevant features when identifying potential biases.

Once potential features were identified, the next step in this analysis stage was understanding the current dataset (i.e., identifying which changes would be needed to facilitate the extraction of the features outlined above). This is presented below.

The current dataset

As Study 2 (Chapter 5) focused on labelling potential ethnic bias in online learning materials, the resulting dataset was not ready for being used in this research cycle. For example, the current dataset had the following attributes: sentences with an explicit reference to an ethnic group (see Section 5.2.2), the excerpts from which each sentence was extracted, and general attributes regarding the origin of the data (i.e., OER’s title; OER’s type; and Discipline). Accordingly, most of the features presented above needed to be computed from the existing data. The steps taken to
perform such feature extraction are described in the subsequent section.

6.2.2 RC4: Design and Construction

The design focused on the strategy used for selecting features from the dataset built in the previous research cycles (Study 2). As shown previously, some features related to the context (e.g., discipline) were already available in the current dataset, while others needed to be extracted. The process of extracting and computing new features based on the existing ones is presented below (feature expansion).

Feature expansion

1. Psycholinguistic markers and language abstraction

These features were derived from both sentences and excerpts. The initial step in computing these features involved extracting the part-of-speech (PoS) tags, a common practice in natural language processing (Hube & Fetahu, 2018; Pryzant et al., 2020). Accordingly, the Stanford PoS Tagger (Toutanova et al., 2003; Toutanvoa & Manning, 2000) available through the Python library “Natural Language Toolkit” (NLTK)\(^2\) was used given its high performance when compared to other tagging models (Goh et al., 2022). Then psycholinguistic and language abstraction features were computed (see Table 6.1 and Table 6.3 respectively).

2. Valence and group mentions

In order to compute valence scores and number of mentions to ethnic groups, the raw text from both sentences and excerpts were used. Furthermore, the valence score was computed based on the Valence Aware Dictionary and sEntiment Reasoner (VADER) tool for sentiment analysis (Hutto & Gilbert, 2014). VADER was chosen based on its popularity and promising performance across a variety of contexts (Hutto & Gilbert, 2014), and given that it is

\(^2\)https://www.nltk.org/
readily available through the NLTK library and has an open license. Accordingly, valence scores were computed for both the sentence and the excerpt, each ranging from -1 to +1. A score of -1 indicates negative language, while a score of +1 denotes positive language. Those features are summarised in Table 6.4.

Table 6.4: List of features to capture group mentions and valence. All features, except for target_ethnic (exclusive to sentences), were computed for both sentences and excerpts.

<table>
<thead>
<tr>
<th>Category</th>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>mentions Ethnic Majority Mentions</td>
<td>ethnic_maj</td>
<td>Number of mentions of White-related ethnic groups</td>
</tr>
<tr>
<td>mentions Ethnic Minority Mentions</td>
<td>ethnic_min</td>
<td>Number of mentions of ethnic minority groups</td>
</tr>
<tr>
<td>mentions Target Group</td>
<td>target_ethnic</td>
<td>= 0 (ethnic minority) or 1 (ethnic majority)</td>
</tr>
<tr>
<td>valence Text Valence</td>
<td>valence</td>
<td>= Vader valence between -1 and +1 (see Section 6.2.2)</td>
</tr>
</tbody>
</table>

3. Context-related features

To conduct the correlation tests, the existing context-related features (i.e., type and discipline of learning materials) from the original dataset needed to be transformed from categorical to numerical values. Accordingly, a one-hot encoding approach was used, which was conducted using the Python library “scikit-learn”. After the transformation, the resulting features were: disc_applied, disc_arts, disc_math, disc_natural, disc_social, type_activ, type_course, type_lesson, and type_other (see Table 6.5).

Table 6.5: Context features converted from categorical to numerical variables.

<table>
<thead>
<tr>
<th>Category</th>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>context</td>
<td>Discipline: Applied Sciences</td>
<td>disc_applied = 1 or 0</td>
</tr>
<tr>
<td>context</td>
<td>Discipline: Arts and Humanities</td>
<td>disc_arts = 1 or 0</td>
</tr>
<tr>
<td>context</td>
<td>Discipline: Mathematics</td>
<td>disc_math = 1 or 0</td>
</tr>
<tr>
<td>context</td>
<td>Discipline: Natural Sciences</td>
<td>disc_natural = 1 or 0</td>
</tr>
<tr>
<td>context</td>
<td>Discipline: Social Sciences</td>
<td>disc_social = 1 or 0</td>
</tr>
<tr>
<td>context</td>
<td>Type: Activity</td>
<td>type_activ = 1 or 0</td>
</tr>
<tr>
<td>context</td>
<td>Type: Course</td>
<td>type_course = 1 or 0</td>
</tr>
<tr>
<td>context</td>
<td>Type: Lesson</td>
<td>type_lesson = 1 or 0</td>
</tr>
<tr>
<td>context</td>
<td>Type: Other</td>
<td>type_other = 1 or 0</td>
</tr>
</tbody>
</table>
To illustrate that encoding approach, consider a sentence extracted from a “lesson” from the discipline “Natural Sciences”. That sentence would have the features type_lesson (i.e., “Type: Lesson”) and disc_natural (i.e., “Discipline: Natural Sciences”) equal to 1 (one) and the other contextual attributes (i.e., category: context) equal to 0 (zero).

4. **Bias score**

The bias score comprised a number between -1 and 1, and was computed based on the number of marks each sentence received:

\[
bias\_score = -1(prop\_no) + 0(prop\_unsure) + 1(prop\_yes) \quad (6.1)
\]

Where,

- \(prop\_no\): proportion of participants that marked the sentence as “No” (i.e., the sentence was considered not biased);

- \(prop\_unsure\): proportion of participants that marked the sentence as “I am not sure” (i.e., they could not decide if the sentence was biased or not); and

- \(prop\_yes\): proportion of participants that marked the sentence as “Yes” (i.e., the sentence was considered biased).

It is important to note that the bias score, while computed based on the overall number of marks each sentence received, represents the perspectives of diverse ethnic groups. This approach, grounded on a sample carefully curated to encompass diverse viewpoints (see Table 5.2 in Chapter 5), aims to create a model that is not confined to the perceptions of a specific group but seeks to find a consensus that is more representative of a broader audience. While this method strives for inclusivity, it has the potential to overlook specific concerns of individual groups, a point that is further reflected upon in
Chapter 7 (Section 7.4.3, “Methodological Implications”).

Figure 6.2 presents all features used in this research cycle and indicates how they were informed by the initial set of features.
Figure 6.2: Overview of the feature expansion process. The green features represent the initial set of features, while the blue features represent the newly created features derived from the initial set. Arrows pointing from green to blue indicate the transformation of the respective green features into the corresponding blue features.
Data analysis and tools

The data analysis process was grounded in the strategy delineated for feature selection, as detailed in Section 3.4. This strategy involved utilising a filter method where a correlation approach was adopted to identify the features significantly related to the bias score, which was defined in the feature expansion section above.

The analysis comprised Pearson’s product-moment correlation tests, a statistical method employed to establish the strength and direction of the association between two continuous variables. To facilitate this analysis, the R Studio software, a premier tool for statistical computing and graphics, was utilised. This environment, integrated with various tools for data analysis, provided a robust platform to compute the correlations efficiently and accurately, leveraging its extensive range of packages and libraries tailored for data science needs.

The outcomes of this analysis are delineated and discussed in the subsequent section, providing insights into the significant features and their relation to the bias score.

6.2.3 RC4: Evaluation and Reflection

After conducting the steps described in the previous sections, it was found that the correlation analysis of ethnic bias and the expanded features indicated a small but statistically significant correlation between the bias score and some features. The features statistically significantly correlated to the bias score were: target_ethnic, disc_arts, self_ref_exc, type_lesson, ethnic_maj, adj_ratio_exc, vrb_ratio_exc, aggressiv_exc, socialisation, emo_stability_exc, unique_word, unique_pos, disc_social, vrb_ratio, aggressiv, and type_activ. Within those features, both positive and negative correlations were found. Accordingly, Table 6.6 presents the full results of the correlation analysis (i.e., Pearson’s Product Moment Correlations), and each column can be interpreted as follows:
• cat.: feature category (i.e., psychological = psycholinguistic markers, abstraction = linguistic abstraction, valence = language valence, mentions = ethnic group mentions, and context = contextual attributes).

• feat.: feature identifier.

• t: the t-statistics of the correlation test.

• df: Degrees of freedom. In this case, the number of comparison groups (c = 2) subtracted from the sample size (n = 345).

• sig.: the test’s p-value. Statistically significant values (i.e., p-value < 0.05) are highlighted.

• cor.: the Pearson Correlation Coefficient (r). Accordingly, values between 0 and 1 indicate a positive correlation, 0 (zero) no correlation, and values between 0 and -1 suggest a negative correlation.

• 95% CI: the 95% confidence interval of r in which lwr indicates the lower bound and upr the upper bound.

The next sections provide more details about the findings and are organised according to each feature category.

Psycholinguistic markers (psychological)

In terms of psycholinguistic markers, the results revealed that self_ref_exc, aggressiv_exc, socialisation, and emo_stability_exc were statistically significantly correlated to the bias score. A negative relationship was found between the Self-Reference Ratio in the excerpt (self_ref_exc) and ethnic bias. This means that, within the sentences considered biased, there was a lower number of self-references (i.e., the text’s authors mentioned themselves less often in those texts).

In contrast, the results showed a positive relationship between ethnic bias and the features: Aggressiveness (aggressiv_exc), Socialisation (socialisation), and Emo-
Table 6.6: Correlations between potential features and ethnic bias. Statistically significant values are highlighted in boldface, and rows are sorted by the Pearson Correlation Coefficient (r).

<table>
<thead>
<tr>
<th>cat.</th>
<th>feat.*</th>
<th>t</th>
<th>df</th>
<th>sig.</th>
<th>cor.</th>
<th>lwr</th>
<th>upr</th>
</tr>
</thead>
<tbody>
<tr>
<td>mentions</td>
<td>target_ethnic</td>
<td>-3.257</td>
<td>343</td>
<td>0.001</td>
<td>-0.173</td>
<td>-0.274</td>
<td>-0.069</td>
</tr>
<tr>
<td>context</td>
<td>disc_arts</td>
<td>-2.976</td>
<td>343</td>
<td>0.003</td>
<td>-0.159</td>
<td>-0.260</td>
<td>-0.054</td>
</tr>
<tr>
<td>psychological</td>
<td>self_ref_exc</td>
<td>-2.957</td>
<td>343</td>
<td>0.003</td>
<td>-0.158</td>
<td>-0.259</td>
<td>-0.053</td>
</tr>
<tr>
<td>context</td>
<td>type_lesson</td>
<td>-2.628</td>
<td>343</td>
<td>0.009</td>
<td>-0.140</td>
<td>-0.242</td>
<td>-0.035</td>
</tr>
<tr>
<td>mentions</td>
<td>ethnic_maj</td>
<td>-2.516</td>
<td>343</td>
<td>0.012</td>
<td>-0.135</td>
<td>-0.237</td>
<td>-0.029</td>
</tr>
<tr>
<td>abstraction</td>
<td>adj_ratio_exc</td>
<td>-2.135</td>
<td>343</td>
<td>0.033</td>
<td>-0.115</td>
<td>-0.217</td>
<td>-0.009</td>
</tr>
<tr>
<td>psychological</td>
<td>self_ref</td>
<td>-1.931</td>
<td>343</td>
<td>0.054</td>
<td>-0.104</td>
<td>-0.207</td>
<td>0.002</td>
</tr>
<tr>
<td>abstraction</td>
<td>adj_ratio</td>
<td>-1.698</td>
<td>343</td>
<td>0.090</td>
<td>-0.091</td>
<td>-0.195</td>
<td>0.014</td>
</tr>
<tr>
<td>abstraction</td>
<td>sent_len</td>
<td>-1.555</td>
<td>343</td>
<td>0.121</td>
<td>-0.084</td>
<td>-0.188</td>
<td>0.022</td>
</tr>
<tr>
<td>context</td>
<td>type_other</td>
<td>-1.265</td>
<td>343</td>
<td>0.207</td>
<td>-0.068</td>
<td>-0.172</td>
<td>0.038</td>
</tr>
<tr>
<td>valence</td>
<td>valence</td>
<td>-0.882</td>
<td>343</td>
<td>0.379</td>
<td>-0.048</td>
<td>-0.152</td>
<td>0.058</td>
</tr>
<tr>
<td>abstraction</td>
<td>pnoun_ratio_exc</td>
<td>-0.660</td>
<td>343</td>
<td>0.509</td>
<td>-0.036</td>
<td>-0.141</td>
<td>0.070</td>
</tr>
<tr>
<td>valence</td>
<td>valence</td>
<td>-0.594</td>
<td>343</td>
<td>0.553</td>
<td>-0.032</td>
<td>-0.137</td>
<td>0.074</td>
</tr>
<tr>
<td>mentions</td>
<td>ethnic_maj_exc</td>
<td>-0.507</td>
<td>343</td>
<td>0.612</td>
<td>-0.027</td>
<td>-0.133</td>
<td>0.078</td>
</tr>
<tr>
<td>abstraction</td>
<td>sent_len_exc</td>
<td>-0.505</td>
<td>343</td>
<td>0.614</td>
<td>-0.027</td>
<td>-0.132</td>
<td>0.079</td>
</tr>
<tr>
<td>abstraction</td>
<td>noun_ratio</td>
<td>-0.285</td>
<td>343</td>
<td>0.776</td>
<td>-0.015</td>
<td>-0.121</td>
<td>0.090</td>
</tr>
<tr>
<td>context</td>
<td>disc_math</td>
<td>-0.132</td>
<td>343</td>
<td>0.895</td>
<td>-0.007</td>
<td>-0.113</td>
<td>0.099</td>
</tr>
<tr>
<td>context</td>
<td>disc_natural</td>
<td>-0.132</td>
<td>343</td>
<td>0.895</td>
<td>-0.007</td>
<td>-0.113</td>
<td>0.099</td>
</tr>
<tr>
<td>context</td>
<td>type_course</td>
<td>-0.113</td>
<td>343</td>
<td>0.910</td>
<td>-0.006</td>
<td>-0.112</td>
<td>0.100</td>
</tr>
<tr>
<td>abstraction</td>
<td>noun_ratio_exc</td>
<td>-0.031</td>
<td>343</td>
<td>0.976</td>
<td>-0.002</td>
<td>-0.107</td>
<td>0.104</td>
</tr>
<tr>
<td>context</td>
<td>disc_applied</td>
<td>0.386</td>
<td>343</td>
<td>0.699</td>
<td>0.021</td>
<td>-0.085</td>
<td>0.126</td>
</tr>
<tr>
<td>mentions</td>
<td>ethnic_min_exc</td>
<td>0.515</td>
<td>343</td>
<td>0.607</td>
<td>0.028</td>
<td>-0.078</td>
<td>0.133</td>
</tr>
<tr>
<td>abstraction</td>
<td>unique_pos_exc</td>
<td>0.568</td>
<td>343</td>
<td>0.571</td>
<td>0.031</td>
<td>-0.075</td>
<td>0.136</td>
</tr>
<tr>
<td>abstraction</td>
<td>adv_ratio_exc</td>
<td>0.758</td>
<td>343</td>
<td>0.449</td>
<td>0.041</td>
<td>-0.065</td>
<td>0.146</td>
</tr>
<tr>
<td>abstraction</td>
<td>pnoun_ratio</td>
<td>0.949</td>
<td>343</td>
<td>0.344</td>
<td>0.051</td>
<td>-0.055</td>
<td>0.156</td>
</tr>
<tr>
<td>abstraction</td>
<td>adv_ratio</td>
<td>0.952</td>
<td>343</td>
<td>0.342</td>
<td>0.051</td>
<td>-0.055</td>
<td>0.156</td>
</tr>
<tr>
<td>psychological</td>
<td>socialisation_exc</td>
<td>1.327</td>
<td>343</td>
<td>0.186</td>
<td>0.071</td>
<td>-0.034</td>
<td>0.176</td>
</tr>
<tr>
<td>psychological</td>
<td>emo_stability</td>
<td>1.382</td>
<td>343</td>
<td>0.168</td>
<td>0.074</td>
<td>-0.031</td>
<td>0.179</td>
</tr>
<tr>
<td>abstraction</td>
<td>unique_word_exc</td>
<td>1.386</td>
<td>343</td>
<td>0.167</td>
<td>0.075</td>
<td>-0.031</td>
<td>0.179</td>
</tr>
<tr>
<td>mentions</td>
<td>ethnic_min</td>
<td>1.677</td>
<td>343</td>
<td>0.094</td>
<td>0.090</td>
<td>-0.016</td>
<td>0.194</td>
</tr>
<tr>
<td>abstraction</td>
<td>vrb_ratio_exc</td>
<td>2.063</td>
<td>343</td>
<td>0.040</td>
<td>0.111</td>
<td>0.005</td>
<td>0.214</td>
</tr>
<tr>
<td>psychological</td>
<td>aggressiv_exc</td>
<td>2.063</td>
<td>343</td>
<td>0.040</td>
<td>0.111</td>
<td>0.005</td>
<td>0.214</td>
</tr>
<tr>
<td>psychological</td>
<td>socialisation</td>
<td>2.521</td>
<td>343</td>
<td>0.012</td>
<td>0.135</td>
<td>0.030</td>
<td>0.237</td>
</tr>
<tr>
<td>psychological</td>
<td>emo_stability_exc</td>
<td>2.780</td>
<td>343</td>
<td>0.006</td>
<td>0.148</td>
<td>0.044</td>
<td>0.250</td>
</tr>
<tr>
<td>abstraction</td>
<td>unique_word</td>
<td>2.807</td>
<td>343</td>
<td>0.005</td>
<td>0.150</td>
<td>0.045</td>
<td>0.251</td>
</tr>
<tr>
<td>abstraction</td>
<td>unique_pos</td>
<td>2.993</td>
<td>343</td>
<td>0.003</td>
<td>0.160</td>
<td>0.055</td>
<td>0.261</td>
</tr>
<tr>
<td>context</td>
<td>disc_social</td>
<td>3.069</td>
<td>343</td>
<td>0.002</td>
<td>0.163</td>
<td>0.059</td>
<td>0.264</td>
</tr>
<tr>
<td>abstraction</td>
<td>vrb_ratio</td>
<td>3.647</td>
<td>343</td>
<td>&lt;0.001</td>
<td>0.193</td>
<td>0.089</td>
<td>0.293</td>
</tr>
<tr>
<td>psychological</td>
<td>aggressiv</td>
<td>3.647</td>
<td>343</td>
<td>&lt;0.001</td>
<td>0.193</td>
<td>0.089</td>
<td>0.293</td>
</tr>
<tr>
<td>context</td>
<td>type_activ</td>
<td>3.854</td>
<td>343</td>
<td>&lt;0.001</td>
<td>0.204</td>
<td>0.100</td>
<td>0.303</td>
</tr>
</tbody>
</table>

*Features ended in "_exc" are the ones extracted from the excerpt (i.e., the text surrounding each sentence).
Emotional Stability – attributes that can indicate the author’s emotional state (see Section 6.2.2) – were statistically significantly correlated to ethnic bias (See Table 6.6).

**Linguistic abstraction** (*abstraction*)

Regarding linguistic abstraction (*abstraction*), four features were statistically significantly correlated to the ethnic bias score: adj_ratio_exc, vrb_ratio_exc, unique_word, unique_pos. On the one hand, the Adjective Ratio within the excerpt (adj_ratio_exc) was negatively correlated to the bias score, which indicates a higher adj_ratio_exc when there is a lower bias score, and a lower adj_ratio_exc when there is a higher bias score. On the other hand, the Verb Ratio within the excerpt (vrb_ratio_exc), Unique Word Ratio (unique_word), and Unique PoS Ratio (unique_pos) were positively correlated to the bias score. As those features can indicate linguistic abstraction (see Section 6.2.2), these findings confirm a potential relationship between linguistic abstraction and bias (Maass, 1999).

**Language valence** (*valence*)

When checking for potential correlations between the valence scores (sentence and context), the findings suggested that none of them was correlated to the bias score. Table 6.6 also depicts those findings in more detail.

**Ethnic group mentions** (*mentions*)

In terms of the features related to ethnic group mentions (*mentions*), two features were statistically significantly correlated to the bias score: Target Group (target_ethnic), and Ethnic Majority Mentions (ethnic_maj). Both presented an inverse relationship between their values and the bias score, which indicates that the presence of terms referring to individuals from a White-related ethnic group was less often considered biased in the sentences.
Contextual attributes (*context*)

In respect of contextual attributes (*context*), the following features showed a statistically significant correlation with ethnic bias: disc_arts, type_lesson, disc_social. In particular, an inverse correlation was found for the discipline “Arts and Humanities” (disc_arts) and when the type of learning material was “Lesson” (type_lesson). That means potential ethnic bias was less likely to appear in that discipline and material type. Conversely, a positive (and statistically significant) correlation was identified between ethnic bias and the discipline “Social Sciences” (disc_social), which suggests that ethnic bias was more likely to be found in learning materials from that discipline.

Key findings and implications for the next research cycle are discussed in the following section.

Key findings and implications

Overall, the tested features had both positive and negative statistically significant correlations with the bias score. For instance, the bias score was negatively correlated to target_ethnic, disc_arts, self_ref_exc, type_lesson, ethnic_maj, and adj_ratio_exc. That suggests that those attributes are more likely to be related to the absence of ethnic bias in the online learning materials analysed. In contrast, vrb_ratio_exc, aggressiv_exc, socialisation, emo_stability_exc, unique_word, unique_pos, disc_social, vrb_ratio, aggressiv, and type_activ were positively associated with the bias score. This suggests that higher values of those features were more likely to be found within the analysed learning materials marked as potentially biased.

In terms of key implications, the findings indicate which aspects of online learning materials are associated with potential ethnic bias. This can inform course designers and online learning platform stakeholders where potential ethnic biases are likely present. For example, the findings suggested a positive correlation between ethnic bias and the discipline of Social Sciences. While that finding could be explained by
frequent discussions about ethnic-related topics (e.g., racism, ethnic identity, etc.) within social sciences, it also suggests the necessity for a more rigorous review of such content before it is presented to students.

In addition, those findings inform the next research cycle in terms of which features could support the automation of bias detection by certain classification models. Accordingly, the features that were statistically significantly correlated to ethnic bias (i.e., target_ethnic, disc_arts, self_ref_exc, type_lesson, ethnic_maj, adj_ratio_exc, vrb_ratio_exc, aggressiv_exc, socialisation, emo_stability_exc, unique_word, unique_pos, disc_social, vrb_ratio, aggressiv, and type_activ) were used as input to different machine learning approaches, which will be discussed next.

6.3 Research Cycle 5

The previous research cycle comprised selecting relevant features that could assist in the identification of ethnic bias in online learning texts. Once those features were selected, this research cycle (RC5) focuses on understanding which LA approaches could be used for classifying those texts as “biased” or “not biased”. As with the other research cycles, RC5 is organised into three main sections: Analysis and Exploration, Design and Construction, and Evaluation and Reflection. Each one is presented below.

6.3.1 RC5: Analysis and Exploration

While several machine learning algorithms have been used to address learning-related issues (see Section 2.7.1), the literature suggests that there is no clear evidence regarding which approach would be suitable for identifying ethnic bias in learning texts. For example, the findings presented in Chapter 4 revealed that existing approaches aiming at bias identification in text were designed for non-education settings and were limited in several aspects, such as not taking contextual elements
into account (more details were provided in Section 2.6). Accordingly, this research cycle focused on answering the following research question:

**RQ3.2: Which classification approaches might be suitable for identifying ethnic bias in text-based online learning materials?**

To answer that question, a range of binary classification algorithms was chosen for implementation in this research cycle. Seven classification algorithms were initially selected: (i) Logistic Regression (LG); (ii) Decision Tree (DT); (iii) Naive Bayes (NB); (iv) Support Vector Machines (SVMs); (v) K-Nearest Neighbours (KNN); (vi) Random Forest (RF); and (vii) Extreme Gradient Boosting (XGB). Those algorithms were selected for different reasons, in particular, because of their popularity for binary classification, promising performance in classification tasks, and the researcher’s familiarity with them. More details about each reason are provided below.

Most of those classifiers (except XGB) were selected given their popularity for binary classification tasks. To illustrate, LG, DT, NB, SVMs, KNN, and RF were used in several studies listed in a systematic literature review that covered binary classification methods (Kumari & Srivastava, 2017). Another systematic literature review (Thangaraj & Sivakami, 2018), which covered text classification techniques, also highlighted the popularity of those algorithms for classification tasks. As the aim of this research cycle was to identify which approach would be suitable for classifying a text as “biased” or “not biased” (i.e., binary classification), those algorithms were chosen as promising candidates.

The latter algorithm, XGB (T. Chen et al., 2015), was selected because of its success in different machine learning tasks. For example, it is well known in the machine learning community for winning several Kaggle competitions (Sagi & Rokach, 2018). In addition, XGB and RF are considered ensemble models, i.e., methods in which multiple machine learning approaches are combined aiming to improve the performance of a single model (Sagi & Rokach, 2018). Accordingly, those
methods were also hypothesised as promising candidates for the bias-identification task in learning texts.

Another (and more pragmatic) reason for choosing those models was the fact the present student had used them in other projects. That would facilitate the implementation of such models in this research cycle (e.g., by reducing the time spent in the coding stage). In addition, implementing models already known would benefit the reliability of the results as it would reduce the chance of potential mistakes during the implementation.

### 6.3.2 RC5: Design and Construction

The design comprised defining the materials and methods used to answer the research question above. The next section provides more details regarding software and tools used for implementing the models indicated above followed by the respective procedures.

#### Materials and Tools

The key materials and tools used during this research cycle comprised the software development environment, libraries, and hardware. Python 3 was used as the primary programming language, given the researcher’s experiences with it and its popularity for implementing machine learning models. In addition, the source code was versioned to a GitHub repository as a way to back up each version, which mitigates potential issues. A version control system also makes it possible to easily share the code with other developers and researchers. In terms of libraries, Scikit-learn (Pedregosa et al., 2011) and Pandas (pandas development team, 2020) were used. The former provides different tools for data analysis and implements core machine learning algorithms. The latter is known for its solid suite of data manipulation tools that are useful for data pre-processing and data transformation. Regarding hardware, an HP EliteBook 820 G3 was used (CPU: 4x 2.3-GHz Intel Core
Once key materials and tools were defined, the implementation was started. The implementation comprised the following steps: (i) defining the evaluation strategy; (ii) preparing the data; (iii) training, testing, and fine-tuning each model; and (iv) stacking (i.e., combining different models). The procedures within each step are detailed below.

**Evaluation and testing strategy**

Before implementing each model, the evaluation and validation strategies were defined. Accordingly, the dataset was initially split into two smaller parts: a training set (80% of the data) and a test set (20% of the data). The training set was used for fine-tuning (i.e., finding the best hyper-parameters) and training the models. The test set was used for validating the models (i.e., for ultimately testing their performance). That approach was adopted in order to understand how the models would perform with new data (i.e., with data not used during the model development).

Another notable step before fine-tuning and training the models was addressing its imbalance aspect as the number of samples labelled “biased” (n=75) was smaller than the number of samples marked “not biased” (n=270). While that does not represent an issue for certain models, some algorithms are sensitive to such an imbalance, which can lead to poor classification accuracy (Chawla et al., 2002; Singhal et al., 2018). To address that issue, a common practice used by the machine learning community is over-sampling the class with fewer entries so that the dataset becomes balanced (i.e., with the same number of samples for each class). Accordingly, the Synthetic Minority Over-sampling Technique (SMOTE) approach was adopted (Chawla et al., 2002), which was chosen for its success across different domains and, therefore, being considered a standard for imbalanced datasets (Fernández et al., 2018). In essence, SMOTE over-samples the minority class by introducing synthetic samples based on the nearest neighbours of the minority class (Chawla et al., 2002).
et al., 2002).

Once the training set was balanced, a Stratified K-Fold Cross-Validation (SKCV) approach was used for fine-tuning and training each model. SKCV is a variation of the traditional K-Fold Cross-Validation (CV) – “a data re-sampling method to assess the generalisation ability of predictive models and to prevent overfitting” (Berrar, 2019, p. 542). In essence, CV uses different subsets of the original data to test and train a model over multiple iteration. While in the traditional CV the subsets are randomly selected (potentially affecting the data distribution), SKCV accounts for the class frequency in the training and validation subset, which preserves the class distribution in each subset. Given those benefits, an SKCV was adopted for fine-tuning and training the models.

To evaluate the performance of the models, the F1-score was chosen as the primary metric due to its ability to harmonise two crucial metrics: “precision” and “recall”, into a single value that reflects the model’s overall ability to identify biased sentences both accurately and comprehensively. Understanding this score is facilitated by first explaining the two concepts it harmonises:

1. **Precision**: This metric reflects the accuracy of the model in identifying biased sentences. It is the ratio of correctly identified “biased” sentences to all the sentences the model flagged as “biased” (including those it mistakenly identified). To put it simply, if one were sorting fruits and wanted to pick out all the apples, the precision would tell how many of the fruits picked were actually apples, helping to avoid mistakenly picking oranges.

2. **Recall**: This metric indicates how comprehensive the model is in finding all biased sentences present. It is the ratio of correctly identified “biased” sentences to all the sentences that are actually “biased” in the material (including those it missed). In the context of the fruit sorting analogy, recall would tell how many of the total apples available were managed to pick, helping to avoid missing out on too many apples.
Accordingly, the F1-score is calculated using the formula:

\[
F1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}
\]  

(6.2)

Beyond the F1-score, weighted averages were also utilised in reporting the strengths of the results, a choice grounded in the intention to counterbalance the dataset’s imbalance. This approach considers the weight of each class based on its size, which is more reflective of the performance across different class sizes. It is important to note that while this method aids in mitigating the effects of data imbalance, it carries its own set of limitations, which are discussed in Section 7.3.2 (under Study 3).

In summary, the F1-score provides a balancing precision and recall based on the positive class. In other words, the precision tells how many sentences are actually biased within all sentences predicted as “biased”, and the recall tells how many predicted sentences are biased within all actually biased sentences. Although the F1-score was chosen for evaluating and testing the models, other metrics were also reported for the final testing: precision, recall, and accuracy. Below, more details about the implementation are provided.

**Data preparation**

This stage comprised transforming the dataset (n = 345 sentences) so that it could be ready for the classification models. The main step during this stage was scaling the data, i.e., adjusting the values of the features so that they lay numerically in the same interval, between zero and one (this transformation is required by some of the algorithms, such as SVMs). Once the data was scaled and ready to be processed, the fine-tuning stage started.
Fine-tuning, training, and testing

Each model was trained individually and tested using the evaluation strategy described above, i.e., SKCV. Initially, each model was trained without changing its default hyper-parameters. Then, they were fine-tuned, which comprised training and evaluating the models with different hyper-parameters in order to find the best performance in terms of the pre-defined score (F1-score).

To find the best hyper-parameters a Random Search strategy was adopted (Bergstra & Bengio, 2012), which was chosen to reduce the time for fine-tuning the models. The random search strategy comprises testing each model against different combinations of hyper-parameters randomly selected from a predefined grid of potential parameters. Then, each combination of randomly chosen values for the hyper-parameters is used to train the model, which is evaluated according to the pre-defined metric (F1-Score). In the end, the combination of hyper-parameter values that led to the best performance is returned. Table 6.7 illustrates the different ranges of hyper-parameters (the “grid”) used as input of the random search.

Table 6.7: Range of hyper-parameters tested for each classifier.

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

* For other hyper-parameters, their default values were kept.
Stacking multiple models

Aiming to find the best models along with the stacking approach, the 7 models were combined in groups of 2, 3, 4, 5, 6 and 7 (all of them). That led to 120 different combinations, which were tested based on the F1-score. Once the best combination was found, considering the F1-score, other metrics were also computed for both learner and final stacking model based on an independent portion of data (test data). The results are presented in the next section, followed by a discussion of key findings.

6.3.3 RC5: Evaluation and Reflection

Fine-tuning and training

After fine-tuning each model individually, a set of hyper-parameters yielded better performance. Table 6.8 illustrates those hyper-parameters of each individual model.

Table 6.8: Hyper-parameters that led to the best F1-score after fine-tuning the models using the Random Search strategy.

<table>
<thead>
<tr>
<th>Model</th>
<th>Hyper-Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>LG</td>
<td>C = 0.452, solver = 'newton-cg'</td>
</tr>
<tr>
<td>SVMs</td>
<td>C = 4.641, gamma = 0.360, probability = True</td>
</tr>
<tr>
<td>NB</td>
<td>var_smoothing = 0.792</td>
</tr>
<tr>
<td>KNN</td>
<td>algorithm = 'ball_tree', leaf_size = 8, n_neighbors = 1, p = 1</td>
</tr>
<tr>
<td>DT</td>
<td>criterion = 'entropy', max_depth = 14, max_features = 7</td>
</tr>
<tr>
<td>RF</td>
<td>max_depth = 48, max_features = 2, min_samples_split = 5, n_estimators = 250</td>
</tr>
<tr>
<td>XGB</td>
<td>objective = 'binary:logistic', booster = 'gbtree', learning_rate = 0.1, n_estimators = 290, max_depth = 14, min_child_weight = 3, gamma = 0.002</td>
</tr>
</tbody>
</table>

* Only showing the hyper-parameters that were manually set, i.e., the values of other parameters were kept default.

Regarding the stacking approach, after testing the 120 possible combinations of the models above, the combination (stack) that led to the best F1-score was the one with the models LG, SVM, NB, KNN, and XGB. In addition, the stacking approach outperformed all models (F1 = 0.88), followed by the SVMs (F1 = 0.86). Accordingly, as the F1-score balances recall and precision, this means the algorithm is accurately identifying instances of bias and non-bias, with few false positives or
false negatives. Those results are also shown in Table 6.9.

Table 6.9: Performance after fine-tuning and stacking models using the training data and the SKCV.

<table>
<thead>
<tr>
<th>Model</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>LG</td>
<td>0.57</td>
<td>0.04</td>
</tr>
<tr>
<td>SVMs</td>
<td>0.86</td>
<td>0.03</td>
</tr>
<tr>
<td>NB</td>
<td>0.63</td>
<td>0.05</td>
</tr>
<tr>
<td>KNN</td>
<td>0.84</td>
<td>0.03</td>
</tr>
<tr>
<td>DT</td>
<td>0.73</td>
<td>0.05</td>
</tr>
<tr>
<td>RF</td>
<td>0.82</td>
<td>0.05</td>
</tr>
<tr>
<td>XGB</td>
<td>0.80</td>
<td>0.04</td>
</tr>
<tr>
<td>STK*</td>
<td>0.88</td>
<td>0.03</td>
</tr>
</tbody>
</table>

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

Models Validation (test data)

Once the models were fine-tuned and trained using the training set, the final step comprised testing their capabilities of classifying ethnic bias in unknown data (i.e., validating them with the test data, which was not used for training). Accordingly, each model was ultimately tested, and their final performance was measured. Those results are shown in Table 6.10.

Key findings and implications

Overall, SVMs presented the best F1-scores during the validation stage (F1 = 0.71), followed by RF (F1 = 0.70). In contrast, NB had the lowest F1-score when submitted to unknown data (F1 = 0.52). This indicates that SVMs and RF models were both reliable in identifying instances of ethnic bias in online learning texts.

In addition, while the stacking approach did not result in the best F1-score during the test, it performed as well as SVMs in terms of best Recall (Recall = 0.75) and best accuracy (Accuracy = 0.75). This means that combining the models LG, SVM, NB, KNN, and XGB could be advantageous when a balanced performance of the three metrics is desirable.
Table 6.10: Models’ final performance after being tested against the test data. “Weighted Avg” refers to the weighted average based on the size of each class, while “Macro Avg” refers to the simple average of the biased and not biased scores.

<table>
<thead>
<tr>
<th>Model</th>
<th>Not Biased</th>
<th>Bias</th>
<th>Macro Avg</th>
<th>Weighted Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>LG</td>
<td>Precision</td>
<td>Recall</td>
<td>F1-score</td>
<td>n</td>
</tr>
<tr>
<td>Not Biased</td>
<td>0.82</td>
<td>0.59</td>
<td>0.69</td>
<td>54</td>
</tr>
<tr>
<td>Biased</td>
<td>0.27</td>
<td>0.53</td>
<td>0.36</td>
<td>15</td>
</tr>
<tr>
<td>Macro Avg</td>
<td>0.55</td>
<td>0.56</td>
<td>0.53</td>
<td>69</td>
</tr>
<tr>
<td>Weighted Avg</td>
<td>0.70</td>
<td>0.58</td>
<td>0.62</td>
<td>69</td>
</tr>
<tr>
<td>SVMs</td>
<td>Not Biased</td>
<td>0.79</td>
<td>0.93</td>
<td>0.85</td>
</tr>
<tr>
<td>Biased</td>
<td>0.33</td>
<td>0.13</td>
<td>0.19</td>
<td>15</td>
</tr>
<tr>
<td>Macro Avg</td>
<td>0.56</td>
<td>0.53</td>
<td>0.52</td>
<td>69</td>
</tr>
<tr>
<td>Weighted Avg</td>
<td>0.69</td>
<td><strong>0.75</strong></td>
<td><strong>0.71</strong></td>
<td>69</td>
</tr>
<tr>
<td>NB</td>
<td>Not Biased</td>
<td>0.88</td>
<td>0.41</td>
<td>0.56</td>
</tr>
<tr>
<td>Biased</td>
<td>0.27</td>
<td>0.80</td>
<td>0.41</td>
<td>15</td>
</tr>
<tr>
<td>Macro Avg</td>
<td>0.58</td>
<td>0.61</td>
<td>0.49</td>
<td>69</td>
</tr>
<tr>
<td>Weighted Avg</td>
<td><strong>0.75</strong></td>
<td>0.49</td>
<td>0.52</td>
<td>69</td>
</tr>
<tr>
<td>KNN</td>
<td>Not Biased</td>
<td>0.79</td>
<td>0.76</td>
<td>0.77</td>
</tr>
<tr>
<td>Biased</td>
<td>0.24</td>
<td>0.27</td>
<td>0.25</td>
<td>15</td>
</tr>
<tr>
<td>Macro Avg</td>
<td>0.52</td>
<td>0.52</td>
<td>0.51</td>
<td>69</td>
</tr>
<tr>
<td>Weighted Avg</td>
<td>0.67</td>
<td>0.65</td>
<td>0.66</td>
<td>69</td>
</tr>
<tr>
<td>DT</td>
<td>Not Biased</td>
<td>0.80</td>
<td>0.59</td>
<td>0.68</td>
</tr>
<tr>
<td>Biased</td>
<td>0.24</td>
<td>0.47</td>
<td>0.32</td>
<td>15</td>
</tr>
<tr>
<td>Macro Avg</td>
<td>0.52</td>
<td>0.53</td>
<td>0.50</td>
<td>69</td>
</tr>
<tr>
<td>Weighted Avg</td>
<td>0.68</td>
<td>0.57</td>
<td>0.60</td>
<td>69</td>
</tr>
<tr>
<td>RF</td>
<td>Not Biased</td>
<td>0.80</td>
<td>0.83</td>
<td>0.82</td>
</tr>
<tr>
<td>Biased</td>
<td>0.31</td>
<td>0.27</td>
<td>0.29</td>
<td>15</td>
</tr>
<tr>
<td>Macro Avg</td>
<td>0.56</td>
<td>0.55</td>
<td>0.56</td>
<td>69</td>
</tr>
<tr>
<td>Weighted Avg</td>
<td>0.70</td>
<td>0.71</td>
<td>0.70</td>
<td>69</td>
</tr>
<tr>
<td>XGB</td>
<td>Not Biased</td>
<td>0.80</td>
<td>0.81</td>
<td>0.81</td>
</tr>
<tr>
<td>Biased</td>
<td>0.29</td>
<td>0.27</td>
<td>0.28</td>
<td>15</td>
</tr>
<tr>
<td>Macro Avg</td>
<td>0.55</td>
<td>0.54</td>
<td>0.55</td>
<td>69</td>
</tr>
<tr>
<td>Weighted Avg</td>
<td>0.69</td>
<td>0.70</td>
<td>0.69</td>
<td>69</td>
</tr>
<tr>
<td>STK*</td>
<td>Not Biased</td>
<td>0.78</td>
<td>0.94</td>
<td>0.86</td>
</tr>
<tr>
<td>Biased</td>
<td>0.25</td>
<td>0.07</td>
<td>0.11</td>
<td>15</td>
</tr>
<tr>
<td>Macro Avg</td>
<td>0.52</td>
<td>0.51</td>
<td>0.49</td>
<td>69</td>
</tr>
<tr>
<td>Weighted Avg</td>
<td>0.67</td>
<td><strong>0.75</strong></td>
<td>0.69</td>
<td>69</td>
</tr>
</tbody>
</table>

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

In contrast, the NB model showed the worst Recall (0.49) and worst Accuracy (0.49) when compared to the other models. In other words, NB may not be as effective in identifying instances of ethnic bias as SVMs and RF (or the stacking approach). However, NB showed the best overall precision (Precision = 0.75) when dealing with unknown data.

Accordingly, future applications of those mechanisms should consider choosing one
of those models according to their respective performance. For example, if one is looking for the model with the highest average precision both for biased and non-biased educational texts, NB emerges as the best option. Similarly, if one is looking for the most accurate model for the same task, SVMs would perform better (or the stack presented above).

Furthermore, the models developed and tested during this research cycle can assist course designers and other stakeholders in preventing potential biases in their online learning materials. For example, some of those models (e.g., SVMs) could be used to pre-screen ethnic biases in online learning platforms so that such biases are removed before the students are exposed to them. This represents a substantial step towards more inclusive and fair educational settings.

Overall, these results provide valuable insights into the optimal machine learning classifiers for detecting ethnic bias in online learning texts. Nonetheless, a notable mention is that these findings are specific to the dataset and methods used in this research cycle, and applications in other contexts should be made with caution.

While those and other implications are discussed in more detail in the next chapter (Chapter 7), next section illustrates how those models can be selected for use in real-world scenarios.

**Recommendations for Educators and Learning Designers**

For the objective of quickly identifying clearly non-biased content in the development phase of online learning materials, the findings support the use of tools grounded in the NB (Naive Bayes) and LG (Logistic Regression) models, given their high precision scores of 0.88 and 0.82 respectively in non-bias detection (see Table 6.10). For example, these models could be advantageous during content reviews preceding the launch of a course, in which they could facilitate a rapid identification of non-biased content, allowing reviewers to focus their attention on a detailed examination of potentially biased materials. Similarly, these two models
could be beneficial in maintaining a bias-free standard by recognising non-biased additions during the content update process, ensuring the continued integrity of the course material.

The findings also suggest that tools grounded in the NB model could be more suitable in scenarios where the emphasis is on flagging potentially biased materials (Recall = 0.80), even at a higher risk of encountering false alarms (Precision = 0.27). This means that such tools could be used for casting a wider range of biased content, making them more likely to encompass all potential instances of bias. Accordingly, NB could benefit various settings distinct from those where precision is prioritised. For example, (i) in environments characterised by a diverse audience and the discussion of sensitive topics, these tools can be deployed to facilitate a meticulous review of all materials, thereby fostering inclusivity and understanding; (ii) during the early stages of content creation, they can be instrumental in identifying and flagging potential biases, contributing to a thorough review process of text-based learning content; (iii) in community-driven platforms, these tools can function as a first line of defence, overseeing contributions and helping to sustain a respectful learning environment.

When the objective is to maintain a balanced approach that navigates between not overlooking many biased contents and avoiding the excessive flagging of non-biased materials, tools based on the SVMs and RF models could be the ideal choice. Their overall F1-scores of 0.71 and 0.70 respectively suggest that they could maintain a reasonable oversight on the content without imposing excessive restrictions, which benefits a balanced review process of learning content. They are particularly beneficial in automated screening processes where platforms aim at balancing between identifying biased and non-biased content, ensuring a streamlined operation. Furthermore, in collaborative settings where multiple stakeholders are involved in content creation, these tools could support a standard of neutrality, facilitating a harmonious and productive content creation process that respects diverse perspectives.
while steering clear of potential biases.

Table 6.11 summarises those aspects and suggests potential scenarios to use (and avoid) each model.

Table 6.11: Model recommendations for bias detection in texts

<table>
<thead>
<tr>
<th>Purpose</th>
<th>Models</th>
<th>When to use</th>
<th>When to avoid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identifying potentially non-biased content (high precision for non-bias)</td>
<td>NB, LG</td>
<td>• Pre-screening content during initial reviews</td>
<td>• Situations requiring a balanced view of both biased and non-biased content</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Quickly identification of non-biased additions during content updates</td>
<td>• Sensitive contexts with a diverse audience</td>
</tr>
<tr>
<td>Thorough flagging of bias, risking false alarms (high recall for bias)</td>
<td>NB</td>
<td>• Reviewing materials in sensitive contexts</td>
<td>• Scenarios where a high precision in identifying non-biased content is the primary focus</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Community moderation in platforms with user-generated content</td>
<td></td>
</tr>
<tr>
<td>Balanced performance (high overall f1)</td>
<td>SVMs, RF</td>
<td>• Automated content screening processes</td>
<td>• Early stages of content creation where a wider net for potential biases is preferable</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Collaborative content creation environments</td>
<td></td>
</tr>
</tbody>
</table>

In all scenarios, it is important to supplement these models with human oversight to ensure a nuanced understanding of the content, blending technology with the depth of human judgement. As discussed in Section 2.4, bias is complex and can depend on factors such as culture and context (Brewer & Yuki, 2007; Chiu et al., 1997; Dovidio & Gaertner, 2010; Hogg, 2016). Therefore, despite the performance of those models, they may still overlook subtle contexts that human reviewers could discern.

6.4 Chapter summary

In this research chapter, two research cycles were presented: RC4 and RC5. Research Cycle 4 comprised selecting potential features that could be used to identify potential ethnic bias in online text-based learning materials. The results indicated that some aspects of language were statistically significantly correlated with ethnic bias, e.g., psycholinguistic features like socialisation and aggressiveness. In addi-
tion, the findings revealed that some context-related features, like ratios related to parts of speech, were also correlated to ethnic bias. Those correlations suggest that these features of language are contributing factors to the emergence of ethnic bias and that interventions targeting such features could be effective in reducing ethnic bias. Drawn from those findings, RC5 focused on training and evaluating seven classification algorithms for identifying ethnic bias in learning texts. The results from RC5 suggested that SVMs yielded the best performance. Overall, SVMs had average performance scores both for biased and non-biased texts of 0.71 (F1-score), 0.75 (accuracy), and 0.69 (precision). This means those models are likely effective for identifying bias in learning texts. Numerous implications can be drawn from those findings. For example, those models can be applied to other texts from online learning platforms to help course designers screen potential biases in online text-based learning materials.

Those models can also be applied to textual datasets and potently spot bias before such being input to machine learning algorithms. In other words, these models have the potential to enhance the quality of training data, consequently contributing to fairer ML algorithms. For instance, the models developed to detect bias in online learning materials can be applied to other textual datasets to identify potential biases, which can help improve the quality of training data and contribute to the development of fairer machine learning algorithms. These models can also assist course designers in screening for biases in online text-based learning materials and help mitigate potential negative impacts on students from marginalised backgrounds. In summary, the use of these models can lead to more equitable learning environments and the development of more ethical and responsible machine learning practices.

A broader discussion of those findings, as well as the ones from the previous research cycles, is presented in the next chapter.
7 | Discussion & Conclusions

7.1 Chapter overview

This doctoral project focused on answering the underlying research question: *How might ethnic bias in text-based online learning materials be automatically identified while considering the subjective nature of bias?* Accordingly, three studies were conducted. Study 1 tested two computational approaches for identifying group bias in 2024 sentences taken from video transcripts from 91 courses across different disciplines. Participants’ responses suggested that these mechanisms were not ready to identify inter-group bias in online learning materials. Drawn from those findings, Study 2 focused on how ethnic bias was perceived by students from different ethnic groups in 345 sentences with explicit references to ethnic groups. Key findings of Study 2 showed that students from ethnic minority and White-related backgrounds perceive ethnic bias differently in learning materials. Then, the data collected in Study 2 were used in Study 3 to identify potential features and classification algorithms for identifying ethnic bias. The findings of Study 3 indicated that psycholinguistic markers, linguistic abstraction, and contextual attributes were statistically significantly correlated with ethnic bias. In addition, SVMs and RF models showed the best performance in identifying instances of ethnic bias in online learning texts, while NB model had the best precision but the worst recall and accuracy.

Accordingly, this chapter covers a general discussion regarding the key findings of this PhD project. Section 7.2 presents key findings, which are organised according to each study. Then, Section 7.3 discusses limitations, which is followed by Section 7.4 highlighting the key theoretical, practical, and methodological implications. Finally, concluding remarks are presented in Section 7.5.
7.2 Key findings

This section covers the key findings of this PhD project and is organised based on each study.

7.2.1 Study 1

During Study 1 of this PhD project, two computational approaches aiming at group bias identification were implemented in online learning materials produced by teachers at The Open University. Like many other institutions, teachers at the OU spend considerable time and effort designing inclusive online learning materials. However, with any human-designed activity, there is the possibility of unintended biases being introduced at the time those materials are created, in particular when those materials are designed on a large scale. Therefore, two computational approaches aiming at identifying bias in text were tested in 2024 sentences taken from video transcripts sampled from 91 courses across several disciplines (Arts, Education, Health, Computing, Natural Sciences, Business, Law, and Social Sciences). Afterwards, crowd-worker students were asked to label a subset of sentences as biased or not biased and their labels were compared to the automated approaches. While both approaches suggested potential biases in the analysed sentences, the responses from participants indicated that there was no major agreement between students on what was actually biased. This suggests that those mechanisms to identify group bias in non-educational contexts are not ready for identifying group bias in learning materials from online learning platforms.

On the one hand, the results indicated that both approaches and most participants agreed that half of the sentences were not biased. On the other hand, the results also indicated there was no major agreement about what sentences were biased. For instance, only three sentences (10%) had 100% of responses concentrated on a single alternative, which means all other sentences had different opinions. In
addition, lower agreement ratios on most sentences do not necessarily imply an individual participant is wrong, but it can suggest that such sentences are (or are not) biased against the social group to which they belong. This concurs with research about inter-group categorisation which has suggested that people tend to favour members of their own group (in-group) rather than members from other groups (out-group) (Grigoryan et al., 2020; Hewstone & Greenland, 2000; Hogg, 2013; Tajfel et al., 1971).

While studying biases within different aspects of linguistics was not new (e.g., Lai and Wilson (2020), and Scaffidi Abbate et al. (2020), findings of Study 1 also suggested that the automation of those aspects to address biases in large educational datasets has not received much attention from the research community. In other words, Study 1 suggested that future research could benefit from bringing together elements of linguistics (e.g., how do group biases manifest in different parts of speech?) and educational technology (e.g., how do we automate the identification of group bias for big educational data?). Those aspects were decisive when designing Study 2 (Chapter 5).

Regarding the approaches implemented in Study 1, Pryzant’s algorithm suggested that nearly 50% of the sentences analysed were most likely to be subjectively biased (see Table 4.3). While their algorithm does not explain why most sentences are classified as biased, their algorithm was designed to identify bias in terms of language subjectivity, i.e., when the language is skewed in terms of opinion, feeling, or taste. However, those aspects do not necessarily harm students, e.g., the sentence “this is perfect” highlighted in Table 4.2 seems acceptable when the word “this” is referring to “keeping good health” but is unacceptable if referring to “Nazism”. Another example is the sentences “all right, boys, we’re going to start off”, and “we are the true Kurdish people” (see Table 4.2).

Accordingly, while those sentences can represent a harmful inclination affecting specific groups (i.e., females and non-Kurdish people can feel excluded from the
environment in which the sentences are present), understanding the actual impact of those potential biases on individuals seems critical before drawing more precise conclusions. Those reflections suggested that an approach to identifying group bias in online learning materials should also consider the actual threat (or benefit) within the respective bias.

In addition, while many studies focused on algorithmic biases (e.g., Kordzadeh and Ghasemaghaei (2021) recently reviewed over 100 articles), the results of Study 1 suggested that the automatic identification of group bias in educational texts had not received much attention from researchers. This assumption arose from the observation that, out of the 84 studies found during the literature search, only two approaches met the selection criteria for implementation. Those findings informed Study 2 in the sense that online learning settings could also benefit from research focusing on the identification of potential biases within online learning materials.

### 7.2.2 Study 2

Drawn from the results of Study 1, Study 2 focused on how ethnic bias was perceived by students from different ethnic groups in text-based learning materials extracted from online platforms. First, a sample of 345 sentences with explicit references to ethnic groups was extracted from Open Educational Resources\(^1\). Then, 193 Higher Education students from different ethnic groups assessed each sentence for potential ethnic bias.

The results suggested bias in more than 20% of the dataset. Furthermore, the results also revealed no consensus on what students perceived as biased, i.e., only 30% of sentences were commonly marked as biased by individuals of both ethnic minority and ethnic majority groups. When participants were asked about the importance of contextual elements that could help them identify ethnic bias in a given 

---

\(^1\)As showed in Section 4.2.3 and Section 7.3.3, a different dataset was used in Study 2 due to limitations in the dataset of Study 1.
sentence (i.e., title, type, discipline, and excerpt, see Section 5.3.2), they expressed the sentences surrounding text (i.e., the excerpt to which a sentence belongs) as the most valuable attribute. Sentences were considered biased for different reasons, e.g., “poor terminology” was the most common indication of ethnic bias. In contrast, “adequate language” was the primary reason for sentences remarked as “not biased”. Below, the findings of Study 2 are highlighted.

In the study, both groups perceived similar levels of ethnic bias (22.3% for ethnic minorities, 25.5% for White students). However, certain sentences were perceived as biased by one group but not the other, as shown by counts 39 and 50 in Figure 5.4. This finding suggests the existence of ethnic bias in online learning materials that were perceived by both ethnic minority populations and individuals identified as White. Although this meets previous investigations regarding the perception of bias across populations (Mellor et al., 2001), these discoveries shed light on the existence of such disparities in online text-based learning materials, in particular OERs. Furthermore, that finding reinforces the importance of recruiting participants from a variety of ethnic backgrounds when investigating ethnic bias in online learning materials, in particular, given that different ethnic groups perceived bias differently.

The results also revealed that students considered sentences either biased or not biased for different reasons. For instance, most of the analysed comments suggested sentences were biased because of the terminology used, i.e., participants considered that the language used to refer to an ethnic group was not appropriate. Although that was the most common reason, it seems the sentences that fall into that category are the least difficult to be identified, as they are related to terms or expressions that can be easily spotted.

In contrast, the sentences identified as biased for other reasons (e.g., generalisation and ethnic segregation) seem more challenging to be identified as such categories seem to require an interpretation of the sentences, and sometimes the context is needed. Similarly, students gave different reasons for the instances they considered
not biased. For example, most comments suggested sentences were not biased because they were not offensive. This suggests that the extent to which certain online learning materials are perceived as offensive or factual depends on who is analysing them. In other words, individuals from certain ethnic backgrounds can feel offended by certain content while others can find the same content adequate.

Results also highlight the importance of contextual elements (e.g., metadata, sentence context) when dealing with ethnic biases. For instance, students indicated the surrounding text of sentences (the ‘excerpt’) as the most important element when assessing bias. This finding has direct implications for potential approaches aiming to identify ethnic biases in online learning materials (both human-based and computer-based). For example, researchers could use such attributes to potentially estimate sentence context, which could mitigate misleading interpretations. Computational approaches could benefit from such attributes by incorporating them into training datasets as a means to tune machine learning predictions. Indeed, other studies have encountered promising results when including contextual attributes in training models (e.g., Nascimento et al. (2018)).

Overall, these findings have both theoretical and practical implications, which are presented in Section 7.4.

### 7.2.3 Study 3

Drawn from the results of Study 2, Study 3 looked at which potential features could be used to identify potential ethnic bias in online text-based learning materials and which classification algorithms could assist the identifying ethnic bias in learning texts. Overall, the results of RC4 indicated that some aspects of language were statistically significantly correlated with ethnic bias, such as psycholinguistic markers, linguistic abstraction, and certain contextual attributes. In contrast, language valence did not show any statistically significant correlation. Additionally, the presence of terms referring to individuals from a white-related ethnic group was less
often considered biased in the sentences. Regarding RC5, which looked at classification algorithms, SVMs and RF models showed the best performance in terms of F1-score, recall, and accuracy. In contrast, the NB model had the worst recall and accuracy compared to the other models while showing the best overall precision. Below, key findings are presented in more detail.

The first part of Study 3 (RC4) explored the relationship between certain features extracted from learning texts (i.e., psycholinguistic markers, linguistic abstraction, language valence, ethnic group mentions, and contextual attributes) and potential ethnic bias. The results suggested that several features were statistically significantly correlated with potential ethnic bias (see Table 6.6). Specifically, the features self_ref_exc, aggressiv_exc, socialisation, and emo_stability_exc in the psycholinguistic markers category were positively correlated with ethnic bias, while adj_ratio_exc was negatively correlated with bias in the linguistic abstraction category. In the ethnic group mentions category, target_ethnic and ethnic_maj were inversely correlated with the bias score. Additionally, context showed that disc_arts and type_lesson were inversely correlated, while disc_social was positively correlated with ethnic bias.

These key findings suggest that certain psycholinguistic markers and contextual attributes are associated with ethnic bias in online learning texts. For example, the use of aggressive language and emotional instability can increase the likelihood of ethnic bias, while the presence of adjectives in text can decrease the likelihood of bias. Additionally, the discipline and type of learning material may also influence the occurrence of ethnic bias in online learning texts, with materials from the social sciences discipline being more likely to contain bias.

The second part of Study 3 (RC5) looked at which commonly used classification models could support the identification of ethnic bias in online learning texts. Accordingly, the results suggested that Support Vector Machines (SVMs) and Random Forest (RF) models had the highest F1-scores in identifying instances of ethnic bias.
in online learning texts. However, the Naive Bayes (NB) model performed poorly in terms of F1-score, recall, and accuracy but had the best precision with unknown data. The stacking approach was found to perform as well as SVMs in terms of recall and accuracy.

This indicates that SVMs and RF models were both reliable in identifying instances of ethnic bias in online learning texts. Similarly, combining the models LG, SVM, NB, KNN, and XGB could also be useful when a balanced performance of the three metrics is desirable. In contrast, NB may not be as effective in identifying instances of ethnic bias as SVMs and RF. However, NB had the best overall precision (Precision = 0.75) when dealing with unknown data.

Overall, numerous implications can be drawn from those findings. For example, the developed models could be applied to other texts from online learning platforms to help course designers screen potential biases in online text-based learning materials. A discussion about the implications of those findings is presented in Section 7.4.

7.3 Limitations

As with any other research, the results of this PhD Project should be cautiously interpreted, and some limitations considered. Indeed, reporting limitations in research can improve the findings’ quality and benefit the interpretation of the evidence presented (Theofanidis & Fountouki, 2018). Accordingly, upon reflecting on different aspects of the present research, several limitations were identified, which are organised below as (i) theoretical limitations; (ii) methodological limitations; (iii) studies design and data limitations; and (iv) research bias and external issues. More details about each one are provided below.
7.3.1 Theoretical limitations

Perhaps the most relevant theoretical limitation is regarding the interdisciplinary nature of this PhD thesis. While bias in online learning is an interdisciplinary subject by nature (social bias plus education plus online technologies), some critics have suggested a few limitations of interdisciplinary research. First, it is argued that interdisciplinary research can make it difficult to establish connections between multiple disciplines, which may lead to a lack of coherence and purpose (Benson & Miller, 1982). Another criticism regarding interdisciplinary research is the assumption it can be harder to be evaluated given that research activities are not organised around a specific discipline (Seeber et al., 2022). To mitigate those limitations, Stokols et al. (2008) suggests emphasising evaluating the research outcomes in terms of academic contributions and real-world applications. Accordingly, while the evaluation stages of each study were presented within the respective chapters, general implications to theory and practice are discussed in Section 7.4.

Another notable theoretical limitation comprises the criticisms the Social Identity Theory has received like the minimisation of individual identity and the way the self-esteem hypothesis was formulated. As indicated in Section 2.4, some aspects of the SIT comprise inter-group comparison and group categorisation, which led to concerns regarding the extent to which such categorisations would reduce individuals’ self-esteem. For instance, such criticisms focus on the potential replication of individualism with a social identity (Hornsey, 2008). Other critics have suggested that the self-esteem hypothesis was not met in full given that it lacks specificity in its own formulation (Rubin & Hewstone, 1998). As indicated in Section 2.4, the self-esteem hypothesis suggests that positive inter-group differentiation elevates self-esteem, and low self-esteem motivates discrimination. To mitigate that limitation, this PhD project incorporated individual differences at its core by seeking the perspectives of particular individuals about ethnic bias, thus potentially making participants more aware of their own unique experiences and viewpoints.
The literature review conducted along with this research also suggests the lack of a robust framework to study bias in text, which directly impacted this PhD project. For instance, the lack of such a framework made it harder to measure bias in text as previous definitions of bias were abstracted, tailored to non-educational contexts, and did not provide an objective way to assess bias. Aiming at overcoming such limitation, Study 2 (Chapter 5) focused on building a baseline of what ethnic bias would look like in text. Once that was established, Study 3 could be started given that training classification models to identify ethnic bias in texts would require a labelled dataset.

Given those limitations, it seems that future research regarding ethnic bias in online learning could benefit from a theoretical framework that bridges different aspects of ethnic bias across different disciplines. In addition, it seems the development of such a framework would benefit from a multidisciplinary research team so that experts from a number of fields could contribute to it.

### 7.3.2 Study design limitations

Other limitations are related to the study design and collected data. In Study 1, for example, two major limitations were the results’ generalisation and limited studies focusing on the identification of bias in text. Regarding Study 2, the major limitations comprised strict ethnic group categories and sample representativity in the qualitative analysis. In Study 3, the key limitations were model generalisation to other fields and limitations of binary classification. Those limitations are organised below according to each study.

**Study 1**

Perhaps, the main design limitation of Study 1 is the result generalisation. For example, Study 1 is limited in terms of the corpus used for analysis, where the number of courses excluded for not having a transcription may prevent generalisations of the
results. For instance, 181 courses were excluded, which represents over 60% of the courses that started in the same period. Therefore, when interpreting the findings, it is important to exercise caution in generalising the results, as contextual factors and unique characteristics of the corpus utilised may influence their applicability to other situations.

Another limitation of Study 1 was the limited number of studies aiming at identifying bias in text-based content. As indicated in Chapter 4, only two studies fit the selection criteria and were ultimately implemented in the context of online learning technologies. While such a limitation may not have a major impact on the findings of this PhD research, the lack of other approaches to identify bias in text prevents further comparisons between the developed approach and those alternatives. Indeed, a common way to measure the performance of new algorithms comprises the comparison of their outputs against the output of previous algorithms. That limitation was mitigated by building the dataset of Study 2, where students were asked to label several sentences in terms of bias. Then, the labels were used to assess the algorithm developed in Study 3 (i.e., the classification model).

While most issues faced in Study 1 were mitigated in Study 2, the scope of this PhD project was delimited to ethnic bias in online text-based learning materials. Therefore, future research regarding ethnic bias in online learning can benefit from considering more varied text corpora. For example, future studies could consider recorded interactions like forum discussions among students and between students and instructors.

**Study 2**

The most noteworthy limitation of Study 2 is the strict recruitment of participants from only four generalised ethnic groups (Asian, Black, Mixed, and White). While that classification is reductionist for not accounting for individuals from more specific ethnic backgrounds (e.g., African American, British Indians, etc.) it was nec-
necessary to mitigate a major sampling issue: most UK and US students identify as White (or related backgrounds). Therefore, without blocking the sample into specific groups the opinion of individuals from ethnic minority groups would be overshadowed by those from a White-related background. In addition, accounting for a more varied number of groups would affect the feasibility of this project in terms of time and budget (more details about those limitations are provided in the next sections).

The qualitative analysis conducted in Study 2 also has some notable limitations. First, the subjective nature of qualitative analysis can be considered subject to researcher bias by some critics, which may affect the findings. That limitation was mitigated by following a rigorous thematic analysis process of identifying and reviewing codes and themes (see Section 5.3). In addition, while the thematic analysis provided potential reasons why students considered each sentence biased (or not biased), any generalisation should be made with caution given the data only represents the opinion of a small portion of UK and US students.

Given those limitations, future research about ethnic bias in online learning texts can benefit from recruiting students from other ethnic groups and countries.

Study 3

In terms of Study 3, a notable limitation is the model generalisation to other contexts. For instance, Study 3 used a dataset built by Higher Education students from the US and UK, therefore, their perspectives regarding what might be considered bias in online learning materials might differ from other populations. In that sense, despite the promising performance of the trained classification models when classifying ethnic bias within the given dataset, it is important to highlight that any generalisations may be limited.

Another limitation pertains to the use of weighted averages in reporting the strengths of the results. While this approach was adopted to counterbalance the dataset’s
imbalance, it might potentially lead to misleading interpretations due to the unevenness of the data and the higher non-bias detection results for most techniques utilised. It is essential to approach the results with a critical lens, considering the potential skew introduced by this method. To offer a more balanced perspective, macro-average scores have been included alongside weighted averages in the results section, facilitating a more robust interpretation of the results.

A further limitation of Study 3 was the type of classification used, i.e., the binary classification. In such a classification approach, only two outputs were assumed when classifying a sentence: either biased or not biased. While it can be argued that some sentences can have different levels of bias, a binary approach was adopted to simplify a complex problem: identifying ethnic bias in online learning materials.

Accordingly, future mechanisms aiming at bias identification in texts could benefit from larger datasets that include not only online learning materials but also user interactions within learning platforms. In addition, projects with a larger scope and budget should also consider assessing more complex algorithms and techniques, for example, advanced language models, and deep learning mechanisms like convolutional neural networks.

### 7.3.3 Other constraints

In addition to the limitations highlighted above, which are directly related to the research process, other constraints are also worth mentioning. Such constraints are related to external issues and are organised below as limited resources, data constraints, and researcher-related constraints.

#### Limited resources

As in any other project, this PhD study was also limited in terms of time and budget. While each step of the research process was planned in advance to accommodate those limitations, some decisions along with this PhD were bounded by such limi-
itations. For example, a larger sample of students could potentially improve the robustness of the data collected, however, a larger budget would also be needed in order to recruit more participants. In addition, more collected data would also require more time to analyse such data, which could compromise other essential activities of a PhD program like training, research dissemination, and thesis writing up within the available (funded) time.

Data constraints

In Study 1, transcriptions from online courses extracted from the Open University Virtual Learning Environment were used to implement existing approaches regarding bias identification. While such a dataset was useful for the scope of this project, it was decided later to use another source of data which could allow the publication of the data afterwards. For instance, some materials were copyrighted such that their publication would not be possible. In addition, some transcriptions could potentially identify instructors, which could affect their privacy if made publicly available. While such limitation did not have a major impact on the aim of this research, the data scope was slightly limited.

Researcher-related constraints

Finally, it is worth mentioning issues that have affected the researcher’s circumstances in conducting this project. Perhaps the most notable constraint was the COVID-19 pandemic, which led to strict restrictions and lockdowns. For instance, all research activities that took place during the lockdown period were conducted from home instead of the university’s campus. While such restrictions did not have a direct impact on the research itself, working away from fellow students have diminished informal, but useful, discussions that could have provided more insights for this project.
7.4 Implications

Despite the limitations discussed above, several implications can be drawn from this doctoral project. These implications were categorised into theoretical (Section 7.4.1), practical (Section 7.4.2), and methodological (Section 7.4.3) in the following sections.

7.4.1 Theoretical implications

The findings of this doctoral project have implications for several fields, including Learning Analytics, Artificial Intelligence, and Social Psychology. In particular, the research conducted along with this PhD journey provides insights into which aspects of learning texts are associated with ethnic bias and how this bias can affect student learning outcomes.

First, Study 2 found that the perception of ethnic bias in learning texts can vary across students of different ethnic groups (see 5.3.3). This insight can guide LA researchers in developing interventions aimed at reducing the effects of bias in students from different backgrounds. For example, interventions tailored to students from particular ethnic groups, who are struggling with their course, can pinpoint the additional support or alternative resources they need to overcome potential language or cultural barriers. Understanding these aspects can also enhance the understanding of performance gaps in online learning platforms highlighted in previous studies (Q. Nguyen et al., 2020; Sabnis et al., 2022).

Second, Study 3 identified features of learning texts (Section 6.2.3) that were statistically significantly associated with ethnic bias. This can also guide future studies in the field of LA. For instance, by incorporating these features into new LA approaches (e.g., predictive models), researchers can gain a more comprehensive understanding of the factors influencing student engagement and performance in the course. As shown in Section 2.7.1, grasping the aspects affecting student behav-
ior is crucial for informing learning design and enhancing their engagement in the course (Dooley & Makasis, 2020; Sedrakyan et al., 2020).

These features can also guide research in Artificial Intelligence towards the design of fairer machine learning models. For instance, since some features were correlated with ethnic bias, they can be instrumental in identifying which text-based datasets contain ethnic bias, leading to the pinpointing of problematic data. This can enhance the quality of a training dataset, fostering the development of fairer machine learning models. As noted in Section 2.5.1, crafting fairer algorithms is a pivotal area of research, especially in reducing algorithmic bias in classification models (R. S. Baker & Hawn, 2021; Mehrabi et al., 2021; Mohamed et al., 2020; Ntoutsi et al., 2020; J. A. Smith, 2015).

Beyond Artificial Intelligence, the field of Social Psychology can also benefit from this doctoral project. For example, the dataset produced in Study 2 – illustrating how students from different ethnic backgrounds perceive inter-group bias in learning texts – can spur new research into social identity and group categorisation in the context of online learning. This is pertinent for Social Psychology as it can shed light on the potential for discrimination and prejudice to occur in online learning environments. As discussed in Section 2.4, Tajfel’s social identity theory posits that individuals define themselves partly through their social group affiliations, which can shape attitudes and behaviors towards members of other groups (Tajfel, 1970; Tajfel et al., 1979). In the context of online learning, students from different ethnic backgrounds can encounter biased or discriminatory content that reinforces negative stereotypes or marginalises their cultural experiences.

Accordingly, by identifying the specific aspects of learning texts that are perceived as biased by students from different ethnic backgrounds, researchers from Social Psychology can gain insights into how social identity and group categorisation operate in online learning environments. This information could be used to develop interventions aimed at reducing the effects of inter-group bias on students, which
would reduce discrimination and prejudice in online learning environments and promote more equitable educational outcomes for all students (Jacoby-Senghor et al., 2016; Mulvey et al., 2022; Skopec et al., 2021).

Overall, the findings of this PhD project have important theoretical implications for improving the quality and fairness of online learning platforms. By identifying the specific aspects of learning texts that are associated with ethnic bias, researchers can develop interventions aimed at reducing the effects of inter-group bias on students, which can lead to less discrimination and prejudice in online learning and promote equitable education for students of different backgrounds.

The next section on “practical implications” builds on this by providing more specific examples of how the findings of this PhD project can inform practice.

### 7.4.2 Practical implications

Perhaps, the most notable practical implication of this PhD study comprises the use of machine learning to support the identification of ethnic bias in online learning texts. For example, the classification models implemented in Study 3 could be incorporated into learning platforms as a screening mechanism for potential biases in large datasets. This could suggest which courses are more susceptible to potential ethnic biases while reducing the time course designers and content authors spend analysing such materials. In essence, this can help stakeholders identify issues related to inter-group bias in educational content, which is essential for promoting equity and inclusion in education (Jacoby-Senghor et al., 2016; Mulvey et al., 2022; Skopec et al., 2021).

Nonetheless, the different machine learning models used in Study 3 have different strengths and weaknesses in terms of accuracy, recall, and precision. This means that choosing a specific model depends on the specific needs and requirements of the task at hand. Accordingly, taking the findings of Study 3 into account, the following choices are recommended:
1. SVMs and RF models are promising when high accuracy and recall rates are required as they performed well in terms of those metrics (F1-scores of 0.71 and 0.70, respectively). In particular, these models can be useful in detecting instances of ethnic bias in online learning texts in cases where false positives could have serious consequences.

2. The stacking approach did not result in the highest F1-score during testing, but it had the best recall and accuracy scores (0.75), which was similar to SVMs. Therefore, the stacking approach could be useful when a balance between accuracy, recall, and precision is required. This could be promising when it is desired to reduce the risk of missing instances of ethnic bias while minimising false positives.

3. In contrast, NB had the lowest Recall (0.49) and Accuracy (0.49), suggesting that it may not be as reliable as other models in correctly identifying instances of ethnic bias. However, when it comes to dealing with unknown data, NB had the best overall precision (0.75). As noted in Section 6.3.2), Precision is the proportion of true positive results out of all positive classifications (instances classified as ethnic bias, whether correctly or incorrectly). This indicates that NB can be useful in situations where it is important to minimise false positives, even if this means sacrificing recall. In essence, NB can be effective in correctly identifying instances of ethnic bias when dealing with new data.

The findings of this PhD project also have practical implications for the design of new curricula. For example, the instances of ethnic bias indicated by students in Study 2 could inform content creators about how ethnic bias looks in online learning materials so that they can avoid such biases in future educational content, such as lessons, activities, textbooks, etc. This could promote the creation of new curricula that also incorporates the feedback of students from different backgrounds regarding what could be considered ethnic bias.
Furthermore, some findings from Study 2 could also support the development of training material for teachers. For example, Section 5.3.3 highlighted some reasons why students considered sentences biased (e.g., offensive language, and generalisations), which again could be incorporated in teaching programs aiming at the development of lecturers’ inclusive skills (e.g., what expressions and terms should be avoided during their classes).

Beyond improving teaching skills, educators could also use this information to develop teaching materials that promote inclusive and respectful language among students. For example, teaching students that certain aspects of language, such as poor terminology, could help them recognise and evaluate bias in language. This could support students in terms of developing awareness about ethnic issues and, therefore, making them more engaged with the topic.

Overall, those findings have practical implications for the design of new curricula, the development of teaching skills, and the creation of teaching materials that promote inclusive and respectful language among students. As discussed in Section 2.5.2, this is important for avoiding ethnic bias in online learning materials, which would promote curriculum decolonisation (Shahjahan et al., 2022; Skopec et al., 2021). In addition, teaching students critical thinking skills to recognise and evaluate ethnic bias in language would promote a more inclusive learning environment and help students become more informed about the topic.

In addition to the practical implications, this PhD project also has important methodological implications, which are presented as follows.

### 7.4.3 Methodological implications

The three studies conducted in this project have important methodological implications that could inform the design of future research in online learning. Those implications are highlighted as follows.
First, Study 1 reinforces the need for designing machine learning mechanisms tailored to specific contexts. For example, the two approaches aiming at identifying bias in non-educational settings marked certain sentences as “biased” that were not considered biased by most students (see Section 4.2.3). This suggests that, when designing classification models, researchers should develop and test approaches that are more sensitive to the specific types of biases that are relevant to educational contexts.

Second, the methods used to list sentences targeting specific groups could be useful for a variety of applications dealing with ethnic content in text. For example, the explicit references to ethnic groups produced during RC2 (see Table 5.1) could be used for screening historical content about ethnic groups. Although that would still require human input, the overall effort would be reduced. Indeed, the approach designed in Study 3 was only possible thanks to a dataset that contained online text-based learning materials related to ethnic groups. In addition, in Study 2, only 112 online learning materials had an explicit reference to an ethnic group (for instance, 112 corresponds to 8% of a total of 1407 online learning materials initially retrieved). This means that looking for this shortlisted set of materials without those explicit references would have taken too much time. In summary, spotting online learning materials with such references is important not only for investigating bias-related issues but for understanding other aspects of ethnic groups in online learning materials (i.e., other ethnic studies).

Third, the design of Study 2 can also inform the design of other studies in educational research. For example, the use of crowd-sourcing and open datasets has been shown effective in reaching diverse samples of learning materials and participants respectively. For example, despite several restrictions due to the COVID-19 pandemic, those mechanisms allowed reaching a diverse sample of students and learning materials. While the advantages of those methods have already been shown in other studies (Chandler et al., 2014; Palmer & Strickland, 2016), this project sug-
gests their potential effectiveness in researching ethnic issues in online learning, demonstrating their usefulness beyond their initial scope and potential for broader applications in research.

Furthermore, the design of Study 2 can also inform the process of labelling data for machine learning algorithms. As mentioned in Chapter 5 (Section 5.2.2), the recruitment of participants was made with caution to promote sample diversity in terms of ethnic groups. After completing Study 2, the findings confirmed that students from different ethnic groups perceived ethnic bias differently (see Section 5.3.3). This means that machine learning models and algorithms used for natural language processing in online learning materials should be designed with sensitivity to potential ethnic bias in mind. In particular, the data labelling process should take into account the diverse perspectives and experiences of different ethnic groups, especially in applications involving variations in language use and diverse cultural contexts.

The core steps taken along with this project can also inform the design of approaches focusing on other inter-group biases (e.g., biases based on gender, social class, religion, etc.). For example, to automate the identification of bias against social classes in learning texts, researchers could consider (1) identifying keywords referring to certain social classes (e.g., “wealthy”, “poor”, “rich”, etc.); (2) filtering learning texts based on those keywords; (3) asking students from different social classes to label those texts; (4) testing different machine learning models based on those labels.

During the later stages of this research, particularly in Study 3, defining and calculating a bias score emerged as a pivotal step, fundamentally influencing the steps carried out in the subsequent stages (i.e., feature selection, and model development). This score (see Section 6.2.2 in Chapter 6) was derived from the collective feedback of a diverse group of participants representing a variety of ethnic backgrounds, a strategy that aimed at fostering a consensus potentially more representa-
7.5 Concluding remarks

This doctoral project looked at the underlying research question: How might ethnic bias in text-based online learning materials be automatically identified while considering the subjective nature of bias? To answer this question, a design-based approach was adopted to guide three studies:

Study 1 utilised pre-existing methods that aimed to detect inter-group bias in text and applied them to online learning materials (RQ1.0: How do computational approaches, that claim to uncover inter-group bias in non-educational texts, perform...
in text-based online learning materials?). Then, Study 2 focused on understanding how ethnic bias manifests in learning texts (RQ2.0: How does bias against certain ethnic groups manifest in text-based online learning materials?). Finally, Study 3 looked at potential features and classification algorithms for automating the identification of ethnic bias in learning texts (RQ3.0: To what extent can Learning Analytics assist in the identification of ethnic bias in text-based online learning materials?).

Despite the limitations presented in Section 7.3 and the challenges along with this PhD journey, the three studies provided key findings that could inform the design of future research in online learning and other fields. Accordingly, some findings and implications are summarised as follows.

Theoretical/Methodological:

1. Study 1 suggested that bias-detection mechanisms should be tailored to specific domains, as certain types of biases are relevant to particular settings. This can inform the design of studies in a range of areas. (Computer Science, Education Technology, Learning Analytics)

2. Crowd-sourcing and open datasets in Study 2 showed promise in researching issues related to ethnic groups in online learning. Those mechanisms can also benefit research on ethnic bias in other areas. (Many)

3. Perception of ethnic bias varied across students of different ethnic groups, which could inform tailored interventions to reduce potential barriers in online learning (Education, Learning Analytics).

4. Certain features of learning texts were associated with ethnic bias, potentially informing future studies in LA to better understand factors influencing student engagement and performance (Learning Analytics).

5. SVMs, RF, and NB models were promising in identifying ethnic bias in learning texts, which could be used to improve the quality of training textual data,

2Potential benefited area(s) indicated between parentheses.
potentially informing the design of fairer machine learning algorithms (Artificial Intelligence).

6. Study 2 provided a dataset with ethnic-related information, which can inform new studies on discrimination in online learning and lead to interventions aimed at reducing the effects of inter-group bias on students (Social Psychology, Educational Psychology).

7. Overall, this project suggested designing machine learning models and algorithms with sensitivity to potential ethnic bias in mind and taking into account the diverse perspectives and experiences of different ethnic groups in the data labelling process. (Artificial Intelligence, Machine Learning)

**Practical:**

8. The classification models could be used to screen courses for potential biases, reducing analysis time for course designers and content authors when integrated into learning platforms. In particular,

   (a) SVMs and RF models performed well in terms of accuracy and recall (F1-scores of 0.71 and 0.70, respectively). This can be useful in detecting ethnic bias in online learning texts when false positives could have serious consequences.

   (b) Stacking approach had the best recall and accuracy scores (0.75), similar to SVMs, but not the highest F1-score during testing. It can be useful to reduce the risk of missing instances of ethnic bias while minimising false positives.

   (c) NB had the best overall precision (0.75) but the lowest Recall (0.49) and Accuracy (0.49). It can be more effective in identifying instances of ethnic bias correctly when dealing with unknown data.

9. The labelled dataset could help content creators avoid ethnic bias in future ed-
ucational content, promoting the creation of inclusive curricula incorporating feedback from students of different backgrounds.

10. Themes indicating potential causes of ethnic bias could be incorporated into teaching programs to develop inclusive skills of lecturers and promote student awareness of ethnic issues.

Overall, this PhD thesis has provided valuable insights into the identification of ethnic bias in online learning materials using machine learning, which could inform the development of interventions to promote inclusiveness and equity. The practical implications of this research include incorporating machine learning models into learning platforms to screen for potential bias and inform course designers and content authors about how to avoid biased material in the future. The methodological implications suggest that researchers should design studies that promote sample diversity and sensitivity to potential biases while exploring the potential of crowdsourcing and open datasets to reach diverse samples of learning materials and participants. Furthermore, the design of machine learning mechanisms and algorithms should consider the diverse perspectives and experiences of different ethnic groups. Ultimately, this research provides an important step towards addressing inequities in online learning platforms and promoting inclusiveness and equity.
References


Armstrong, M., Dopp, C., & Welsh, J. (2022). Design-Based Research: What is DBR, why might one do it, and how does one do it well? In *Education research* (pp. 1–6). https://edtechbooks.org/education_research/design_based_research


Bolarinwa, O. (2015). Principles and methods of validity and reliability testing of questionnaires used in social and health science researches. *Nigerian Post-


Cernusca, D. (2007). *A design-based research approach to the implementation and examination of a cognitive flexibility hypertext in a large undergraduate course* [Doctoral dissertation, University of Missouri-Columbia].


du Boulay, B., Poulouvasilis, A., Holmes, W., & Mavrikis, M. (2018). What does the research say about how Artificial Intelligence and Big Data can close the achievement gap. Enhancing learning and teaching with technology, 316–327.


of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL’05), 363–370.


Holmberg, J. (2019). Designing for added pedagogical value: A design-based research study of teachers’ educational design with ICT [Doctoral dissertation, Department of Computer and Systems Sciences, Stockholm University].


İnan-Kaya, G., & Rubie-Davies, C. M. (2022). Teacher classroom interactions and behaviours: Indications of bias. Learning and Instruction, 78, 101516.


OER Commons & Open Education. (2007). https://www.oercommons.org/about


Pearson, K. (1900). X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302), 157–175.


A | Literature Review Protocols

A.1 General protocol

<table>
<thead>
<tr>
<th>Question</th>
<th>Search String</th>
</tr>
</thead>
<tbody>
<tr>
<td>What are the advantages and difficulties associated with online learning, and how do these factors influence the experiences and outcomes of students? (Section 2.3)</td>
<td>(&quot;affordances&quot; OR &quot;advantages&quot; OR &quot;benefits&quot; OR &quot;challenges&quot; OR &quot;barriers&quot; OR &quot;difficulties&quot;) AND (&quot;online learning&quot; OR &quot;e-learning&quot; OR &quot;distance learning&quot; OR &quot;remote learning&quot;) AND (&quot;impact&quot; OR &quot;influence&quot; OR &quot;effects&quot;) AND (&quot;students&quot; OR &quot;learners&quot;)</td>
</tr>
<tr>
<td>What are the fundamentals of inter-group bias and how does it inform other social issues like prejudice and stereotypes? (Section 2.4)</td>
<td>(&quot;fundamentals&quot; OR &quot;concept&quot; OR &quot;principle&quot; OR &quot;theory&quot;) AND (&quot;inter-group bias&quot; OR &quot;intergroup bias&quot; OR &quot;out-group bias&quot; OR &quot;in-group bias&quot;) AND (&quot;social issues&quot; OR &quot;prejudice&quot; OR &quot;stereotypes&quot; OR &quot;influence&quot; OR &quot;impact&quot;)</td>
</tr>
<tr>
<td>How can bias be incorporated into online learning and what impact does it have on students? (Section 2.5)</td>
<td>(&quot;inter-group bias&quot; OR &quot;stereotype&quot; OR &quot;implicit bias&quot; OR &quot;algorithmic bias&quot; OR &quot;gender bias&quot; OR &quot;racial bias&quot; OR &quot;ethnic bias&quot; OR &quot;socioeconomic bias&quot; OR &quot;disability bias&quot;) AND (&quot;online learning&quot; OR &quot;e-learning&quot; OR &quot;virtual learning environment&quot; OR &quot;distance education&quot; OR &quot;technology-enhanced learning&quot;) AND (&quot;implications&quot; OR &quot;impact&quot; OR &quot;effects&quot; OR &quot;influence&quot;) AND (&quot;student engagement&quot; OR &quot;interaction&quot; OR &quot;learning outcomes&quot; OR &quot;academic performance&quot;)</td>
</tr>
<tr>
<td>How can learning analytics be used to address subjective and complex issues in online learning environments? (Section 2.7)</td>
<td>(&quot;learning analytics&quot; OR &quot;educational data mining&quot; OR &quot;data-driven education&quot;) AND (&quot;complex&quot; OR &quot;complex issues&quot; OR &quot;complex problems&quot; OR &quot;complex challenges&quot;) AND (&quot;solution&quot; OR &quot;solving&quot; OR &quot;addressing&quot; OR &quot;tackling&quot; OR &quot;resolving&quot;)</td>
</tr>
</tbody>
</table>
A.2 Approaches for identifying bias in texts

<table>
<thead>
<tr>
<th><strong>Objective:</strong></th>
<th>To uncover existing computational approaches that identify inter-group bias in text</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Population:</strong></td>
<td>Stakeholders interested in detecting inter-group bias in text</td>
</tr>
<tr>
<td><strong>Intervention:</strong></td>
<td>Computational techniques for identifying inter-group bias in text</td>
</tr>
<tr>
<td><strong>Comparison:</strong></td>
<td>Effectiveness of different computational approaches or comparison with manual methods</td>
</tr>
<tr>
<td><strong>Outcome:</strong></td>
<td>Successful detection of inter-group bias using computational methods</td>
</tr>
<tr>
<td><strong>Context:</strong></td>
<td>Various domains in which inter-group bias may appear in text (e.g., online platforms)</td>
</tr>
</tbody>
</table>

**RQ1.0:** Which computational approaches might identify inter-group bias in text?

**Keywords (synonyms)**
- Inter-group (intergroup, in-group, out-group)
- Bias (“bias in text”, “bias in sentence”)
- Detection (identification, analysis, measurement)
- Computational approach (natural language processing, machine learning, text mining, algorithm)

**Search string**
- (“inter-group” OR intergroup OR “in-group” OR “out-group”) AND (bias OR “bias in text” OR “bias in sentence”) AND (detection OR identification OR analysis OR measurement) AND (“computational approach” OR “natural language processing” OR “machine learning” OR “text mining” OR algorithm)

**Sources**
- Google Scholar (http://scholar.google.com/)
- IEEE Digital Library (http://ieeexplore.ieee.org)
- ISI Web of Science (http://www.isiknowledge.com)
- ScienceDirect (http://www.sciencedirect.com)
- Scopus (http://www.scopus.com)
- Springer Link (http://link.springer.com)
- Mendeley (personal database)

**Inclusion criteria**
- Empirical and primary studies
- Peer-reviewed (journals or conferences)
- Recent studies (2010-2020)
- Reproducible approach
- Approaches to identify inter-group bias in text

**Exclusion criteria**
- Not empirical or secondary studies (e.g., surveys)
- Not peer reviewed
- Published before 2010
- Manual or not automated approach
- Studies not identifying bias in text
- Studies not related to inter-group bias
- Studies not fully available in English

**Quality Assessment Checklist**
- Possible answers: Yes (1.0); Can’t tell (0.5); or No (0.0)

1. Did the trial address a clearly focused issue?
2. Was the assignment of patients to treatments randomized?
3. Were all the patients who entered the trial properly accounted for at its conclusion?
4. Were patients, health workers and study personnel ‘blind’ to treatment?
5. Were the groups similar at the start of the trial?
6. Aside from the experimental intervention, were the groups treated equally?
7. Was the treatment effect large enough?
8. Was the estimate of the treatment effect precise?
9. Can the results be applied to the local population, or in your context?
10. Were all clinically important outcomes considered?
11. Are the benefits worth the harms and costs?
Dear Josmario,

This message confirms that the research protocol for the following research project, as submitted for ethics review, has been given a favourable opinion on behalf of The Open University Human Research Ethics Committee.

Project title: Uncovering Subjective Bias in Virtual Learning Con(texts)

HREC approval date: 18/06/2020

As part of your favourable opinion, it is essential that you are aware of and comply with the following:

1. In light of the current exceptional circumstances relating to the COVID-19 pandemic, you are not permitted to begin data collection which requires any face-to-face interactions with participants until further notice. If your study does involve face-to-face interactions with participants, you will receive official notification from HREC when data collection is permitted to commence.

2. You are responsible for notifying the HREC immediately of any information received by you, or of which you become aware which would cast doubt on, or alter, information in your original application, in order to ensure your continued safety and the good conduct of the research.

3. It is essential that you contact the HREC with any proposed amendments to your research, for example - a change in location or participants. HREC agreement needs to be in place before any changes are implemented, except only in cases of emergency when the welfare of the participant or researcher is or may be affected.

4. Your HREC reference number has to be included in any publicity or correspondence related to your research, e.g. when seeking participants or advertising your research, so it is clear that it has been agreed by the HREC and adheres to OU ethics review processes.

5. Researchers should have discussed any project-related risks with their Line Manager and/or Supervisor, to ensure that all the relevant checks have been made and permissions are in place, prior to a project commencing, for example compliance with IT security and Data protection regulations.

6. Researchers need to have read and adhere to relevant OU policies and guidance, in particular the Ethics Principles for Research with Human Participants and the Code of Practice for Research - [http://www.open.ac.uk/research/governance/policies](http://www.open.ac.uk/research/governance/policies)

7. The Open University's research ethics review procedures are fully compliant with the majority of research council, professional organisations and grant awarding bodies research ethics guidelines. Where required, this message is evidence of OU HREC support and can be included in an external research ethics review application. The HREC should be sent a copy of any external applications, and their outcome, so we have a full ethics review record.

B | Ethics

B.1 Ethics Approval

HREC/3593/Albuquerque: HREC Favourable Opinion

Research-REC-Review <research-rec-review@open.ac.uk>

Wed 24/06/2020 11:59

To: Josmario.Albuquerque <josmario.albuquerque@open.ac.uk>; Research-REC-Review <research-rec-review@open.ac.uk>

Cc: Bart.Rienties <bart.rienties@open.ac.uk>

Dear Josmario,

This message confirms that the research protocol for the following research project, as submitted for ethics review, has been given a favourable opinion on behalf of The Open University Human Research Ethics Committee.

Project title: Uncovering Subjective Bias in Virtual Learning Con(texts)

HREC approval date: 18/06/2020

As part of your favourable opinion, it is essential that you are aware of and comply with the following:

1. In light of the current exceptional circumstances relating to the COVID-19 pandemic, you are not permitted to begin data collection which requires any face-to-face interactions with participants until further notice. If your study does involve face-to-face interactions with participants, you will receive official notification from HREC when data collection is permitted to commence.

2. You are responsible for notifying the HREC immediately of any information received by you, or of which you become aware which would cast doubt on, or alter, information in your original application, in order to ensure your continued safety and the good conduct of the research.

3. It is essential that you contact the HREC with any proposed amendments to your research, for example - a change in location or participants. HREC agreement needs to be in place before any changes are implemented, except only in cases of emergency when the welfare of the participant or researcher is or may be affected.

4. Your HREC reference number has to be included in any publicity or correspondence related to your research, e.g. when seeking participants or advertising your research, so it is clear that it has been agreed by the HREC and adheres to OU ethics review processes.

5. Researchers should have discussed any project-related risks with their Line Manager and/or Supervisor, to ensure that all the relevant checks have been made and permissions are in place, prior to a project commencing, for example compliance with IT security and Data protection regulations.

6. Researchers need to have read and adhere to relevant OU policies and guidance, in particular the Ethics Principles for Research with Human Participants and the Code of Practice for Research - [http://www.open.ac.uk/research/governance/policies](http://www.open.ac.uk/research/governance/policies)

7. The Open University's research ethics review procedures are fully compliant with the majority of research council, professional organisations and grant awarding bodies research ethics guidelines. Where required, this message is evidence of OU HREC support and can be included in an external research ethics review application. The HREC should be sent a copy of any external applications, and their outcome, so we have a full ethics review record.
8. At the end of your project you are required to assess your research for ethics related issues and/or any major changes. Where these have occurred you will need to provide the Committee with a HREC final report to reflect how these were dealt with using the template on the research ethics website - http://www.open.ac.uk/research/governance/ethics/human/review-process/final-report (HREC Final Report form)

Sent on behalf of the Human Research Ethics Committee

Dr Claire Hewson                  Professor Louise Westmarland                  Dr Duncan Banks
Chair                                       Deputy Chair                                                   Deputy Chair

Human Research Ethics Committee - Research, Enterprise and Scholarship (RES)
The Open University, Walton Hall, Milton Keynes, MK7 6AA

Email: research-rec-review@open.ac.uk       Tel: 01908 654849
http://www.open.ac.uk/research/governance/ethics/human

This email is private, confidential, and intended for the named recipient. It does not necessarily reflect the views of the Open University or any of its subsidiaries. If you are not the intended recipient or have received this email in error please note that it may be subject to copyright and/or legal privilege. Any disclosure, copying, distribution or reliance thereon is strictly prohibited. If you have received it in error, please contact the sender immediately and confirm that it has been deleted from your system and any hard copies destroyed.

Please consider the environment before printing this email.
B.2 Information Security

CERTIFICATE of ACHIEVEMENT

This is to certify that

Josmario Albuquerque

has completed the course

Information Security Awareness module

February 24, 2020
B.3 GDPR

CERTIFICATE of ACHIEVEMENT

This is to certify that

Josmario Albuquerque

has completed the course

GDPR

January 6, 2020

The Open University
C | Online task

C.1 Participant Information Sheet

<table>
<thead>
<tr>
<th>Participant Information Sheet</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>About the study</strong></td>
</tr>
<tr>
<td>The aim of this study is to understand how different racial biases manifest in educational resources from online platforms.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>What will I be asked to do?</th>
</tr>
</thead>
<tbody>
<tr>
<td>You will be shown excerpts of different texts extracted from a range of learning materials collected from open educational resources from across the globe like activities, lesson plans, textbooks, etc. Then, you will be asked to decide whether those excerpts are biased against an ethnic group. The whole study shall last approximately 15-20 minutes. Detailed instructions will be given in the next page.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Are there any potential risks in taking part?</th>
</tr>
</thead>
<tbody>
<tr>
<td>There is no significant risk associated with this study. Nonetheless, some excerpts may contain terms related to ethnic groups that may cause some discomfort. If at any moment you feel uncomfortable, you have the right to interrupt the experiment immediately.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Are there any benefits in taking part?</th>
</tr>
</thead>
<tbody>
<tr>
<td>You will receive £2.00 in exchange for your participation when you complete the study. The payment will take place via Prolific once you complete the survey.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>What happens to the data I provide?</th>
</tr>
</thead>
<tbody>
<tr>
<td>By the end of the study, you shall be asked your country of residence, ethnic group, gender, current level of education, and year of birth. This will be processed in accordance with the General Data Protection Regulation (GDPR). For instance, the findings of this study may be published in journals and conferences, but your responses will not be shown isolated, i.e., your responses will be combined with the ones from other participants. In addition, your personally identifiable information (except for the Prolific ID) will not be recorded during this study so that your data is inherently anonymous. If you chose to contact the researcher via your personal email service (not via Prolific), the researcher will be able to see your email address. In any case, your information will be treated as strictly confidential.</td>
</tr>
</tbody>
</table>
Who has reviewed this research project?
This research project has been reviewed by, and received a favourable opinion from, The Open University Human Research Ethics Committee - HREC reference number: HREC/3593/Albuquerque.

Who do I contact if I have queries about this study?
If you would like to discuss the research with someone, please contact the lead researcher, Josmario Albuquerque (jas2665@open.ac.uk). You may also use the Prolific platform to prevent your email address being disclosed to the researcher. If you have any other concerns, please contact Prof Bart Rienties (bart.rienties@open.ac.uk).

Your rights.
You have the right to withdraw from this project at any time without giving reasons and without consequences to you. If you withdraw from the study, your data will automatically be removed from all analyses. If you do finish the study, you can still ask to withdraw without being identified - ask the researcher via the Prolific platform or via email as indicated above.

C.2 Electronic Consent Form

**Electronic Consent Form**

1. I understand that my participation is voluntary, and that I can withdraw from the study at any time (by closing the browser).
2. I have read and understood the participant information sheet above.
3. I consent to the recording of my responses to the study, my country of residence, my ethnic group, my preferred gender identity, my current level of education, and my year of birth.
4. I am 18 years old or over
5. I understand that by completing this survey, I give my consent to take part in this study.
C.3 Instructions

1. You will be shown 15 sentences extracted from different learning materials (e.g., lab activity, lesson plan, textbooks, etc.). Your task is to classify each sentence as biased or not biased.

2. A sentence is biased when it is inclined against (negative bias) or in favour of (positive bias) a group or individual. For this study, you shall focus on negative biases that can affect individuals based on their ethnic group.

3. Each sentence contains a term/expression related to an ethnic group, for example, “black student”, “white woman”, etc. When analysing the excerpt, consider whether the individuals from that ethnic group would feel excluded or underrepresented based on the language used.

4. Bias in language can be very objective and explicit: you may identify it based on inappropriate words or expressions. But, it can also be subjective and implicit: you may need to consider the sentence meaning and context. With that in mind, we have provided some details of the learning material from which a sentence was extracted. When analysing a sentence, consider those contextual clues, and use your best judgement.

5. The number of biased sentences in this study is unknown. It could be as low as 0 and as high as 15.

6. You will be invited to comment on your responses if you wish.

7. Please, answer each question honestly. Attention checks have been added along with the task, and your responses will not be considered if we realise you made arbitrary choices. The keyword for the next page is “bias”.
APPENDIX C. ONLINE TASK

C.4 QUESTIONS

C.4 Questions

C.4.1 Labelling question - template

Sentence 1: “The lines between refined white womanhood and degraded enslaved Black femaleness were no longer so clearly defined.”

This sentence was extracted from the following learning material:

Material type: Lesson
Discipline: U.S. History
Title: Reconstruction

Excerpt: “(...)Advocates for women’s suffrage were largely confined to the North, but southern women were experiencing social transformations as well. [...] The lines between refined white womanhood and degraded enslaved Black femaleness were no longer so clearly defined.] Moreover, during the war, southern white women had been called on to do traditional men’s work, chopping wood and managing businesses. While white southern women decided whether and how to return to their prior status, African American women embraced new freedoms and a redefinition of womanhood(...)”

1. Is this sentence biased against the individual(s) referred to as “white woman”?
   - Yes, I am absolutely certain it is biased
   - Yes, I am mostly certain it is biased
   - Yes, I am somewhat certain it is biased
   - I am not sure
   - No, I am somewhat certain it is NOT biased
   - No, I am mostly certain it is NOT biased
   - No, I am absolutely certain it is NOT biased

Please, briefly explain the reason for your answer:

C.4.2 Context-related questions
Based on the previous questions, please indicate the importance of the following elements when deciding if the sentences were biased or not.

<table>
<thead>
<tr>
<th>The material’s type (e.g., lesson plan, book, activity)</th>
<th>1. Not important</th>
<th>2.</th>
<th>3.</th>
<th>4.</th>
<th>5. Very important</th>
</tr>
</thead>
<tbody>
<tr>
<td>The discipline (e.g., History, Law, etc.)</td>
<td>1. Not important</td>
<td>2.</td>
<td>3.</td>
<td>4.</td>
<td>5. Very important</td>
</tr>
<tr>
<td>The material’s title</td>
<td>1. Not important</td>
<td>2.</td>
<td>3.</td>
<td>4.</td>
<td>5. Very important</td>
</tr>
<tr>
<td>The text excerpt with the sentence</td>
<td>1. Not important</td>
<td>2.</td>
<td>3.</td>
<td>4.</td>
<td>5. Very important</td>
</tr>
</tbody>
</table>
D | Thematic analysis

The table below delineates the steps from the individual sentences judged by participants for potential bias against a specified ethnic group (Keyword) to the final themes identified through thematic analysis. In particular, the columns represent:

- **Sentence**: the sentence extracted from a learning material and assessed for potential ethnic bias.
- **Keyword**: the keyword within the sentence associated to the ethnic group referred to.
- **Label**: computed based on the majority of marks the sentence received (yes = biased; no = not biased).
- **Comment**: participant comment justifying their judgments (i.e., why they considered the sentence biased or not biased).
- **Code**: the code derived during the thematic analysis of participants’ comments.
- **Theme**: the final Theme after the coding process.
## Thematic analysis results.

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Keyword</th>
<th>Label</th>
<th>Comment</th>
<th>Code</th>
<th>Theme</th>
</tr>
</thead>
<tbody>
<tr>
<td>This was Malcolm X's curt explanation of why he did not favor integration of blacks with whites in the United States</td>
<td>whites</td>
<td>yes</td>
<td>It talks about a “white nation” and how black people need to create a nation just for themselves because, in their view, the white nation is corrupt. The text makes it clear that the speaker is against the integration of blacks and whites.</td>
<td>ethnic segregation</td>
<td>ethnic preference or segregation</td>
</tr>
<tr>
<td>The lines between refined white womanhood and degraded enslaved Black femaleness were no longer so clearly defined</td>
<td>white</td>
<td>yes</td>
<td>The sentence implies white women cannot be underprivileged in any way and “white” is not capitalized like “Black”.</td>
<td>ethnic/race superiority</td>
<td>ethnic preference or segregation</td>
</tr>
<tr>
<td>We hold as undeniable truths that the governments of the various States and of the confederacy itself were established exclusively by the white race for themselves and their posterity that the African race had no agency in their establishment that they were rightfully held and regarded as an inferior and dependent race and in that condition only could their existence in this country be rendered beneficial or tolerable</td>
<td>white</td>
<td>yes</td>
<td>Yes it seems like the “white race” in this context is viewed as better or more superior in this particular context.</td>
<td>ethnic/race superiority</td>
<td>ethnic preference or segregation</td>
</tr>
<tr>
<td>I might start with the college president A thought I have fromt he viewpoint of Malcolm McLean is I am thrilled that we have these black women working in such an important job at Langley</td>
<td>black</td>
<td>yes</td>
<td>It sounds surprised that black women could be as smart as white women.</td>
<td>ethnic/race superiority</td>
<td>ethnic preference or segregation</td>
</tr>
<tr>
<td>In this lesson students discuss the general trend in the first half of the 19th century to extend the right to vote to more white males</td>
<td>white</td>
<td>yes</td>
<td>again, they are paving a way for white males</td>
<td>inappropriate promotion/credit</td>
<td>ethnic preference or segregation</td>
</tr>
</tbody>
</table>
### Thematic analysis results.

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Keyword</th>
<th>Label</th>
<th>Comment</th>
<th>Code</th>
<th>Theme</th>
</tr>
</thead>
<tbody>
<tr>
<td>Four hundred black and white students associated with CORE Congress of Racial Equality and SNCC Student Nonviolent Coordinating Committee planned to ride Greyhound buses through Alabama Louisiana North Carolina Virginia and other states to protest illegal segregation in the South</td>
<td>white</td>
<td>yes</td>
<td>race is included in the sentence o credit white people for doing the same thing many black people dedicate their lives to because it is their lived experience, not experienced by white people.</td>
<td>inappropriate</td>
<td>promotion/credit</td>
</tr>
<tr>
<td>Many free blacks were able to become businessmen and leaders</td>
<td>blacks</td>
<td>yes</td>
<td>It is a sweeping generalisation which I find derogatory</td>
<td>inappropriate</td>
<td>generalisation</td>
</tr>
<tr>
<td>On December 22 1871 R Latham of Yorkville South Carolina wrote to the New York Tribune voicing the beliefs of many white southerners as he declared that &quot;the same principle that prompted the white men at Boston disguised as Indians to board during the darkness of night a vessel with tea and throw her cargo into the Bay clothed some of our people in Ku Klux gowns and sent them out on missions technically illegal</td>
<td>white</td>
<td>yes</td>
<td>This could be perceived as including all white men with the ridiculous nonsense of the ku klux klan. It is not worded very well and could be considered biased.</td>
<td>inappropriate</td>
<td>generalisation</td>
</tr>
<tr>
<td>In black families is the mother always dominant</td>
<td>black</td>
<td>yes</td>
<td>It implies a generalisation which is unhelpful and smacks of racism</td>
<td>inappropriate</td>
<td>generalisation</td>
</tr>
<tr>
<td>Therefore not only were more white males allowed to vote but that vote also had a direct effect on the outcome of presidential elections</td>
<td>white</td>
<td>yes</td>
<td>This sounds a lot like blaming history on white males as a whole, rather than the white males in power causing the chaos.</td>
<td>inappropriate</td>
<td>generalisation</td>
</tr>
</tbody>
</table>
Thematic analysis results.

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Keyword</th>
<th>Label</th>
<th>Comment</th>
<th>Code</th>
<th>Theme</th>
</tr>
</thead>
<tbody>
<tr>
<td>Call attention to the story of the slave Randall that closes Browns first chapter pages 1619 in the electronic text and to episodes in which Brown himself offers resistance such as his snowball battle with a group of white boys Chapter 3 page 28 and his repeated attempts to escape Chapter 2 page 21 Chapter 7 pages 6568 and Chapter 10 pages 8993</td>
<td>white boy</td>
<td>yes</td>
<td>The text implies that Brown’s conflict with the white boys came about as a result of slavery, but still is biased against the white boys for assuming that they are racist even if is true that that is most likely the case.</td>
<td>inappropriate generalisation</td>
<td>generalisation</td>
</tr>
<tr>
<td>Women like Nannie Helen Burroughs and Virginia Broughton leaders of the Baptist Woman’s Convention worked to protect Black women from sexual violence from white men</td>
<td>white men</td>
<td>yes</td>
<td>Suggests all white men partake in rape</td>
<td>inappropriate generalisation</td>
<td>generalisation</td>
</tr>
<tr>
<td>This was Malcolm Xs curt explanation of why he did not favor integration of blacks with whites in the United States</td>
<td>whites</td>
<td>yes</td>
<td>It’s someone quite strong opinion</td>
<td>inappropriate tone/connotation</td>
<td>misleading statement</td>
</tr>
<tr>
<td>No one knows how many temporarily lost the vote but one estimate was that it was as high as 10000 to 15000 out of a total white population of roughly eight million</td>
<td>white population</td>
<td>yes</td>
<td>Looked like people did not want to vote</td>
<td>inappropriate tone/connotation</td>
<td>misleading statement</td>
</tr>
<tr>
<td>The fact the college president invites black people to his party</td>
<td>black people</td>
<td>yes</td>
<td>The context in which it is written talks about the negative racial climate experienced by black people at the time. The way it is written suggests that black people present at the president’s party is a novelty and not one that is happily accepted</td>
<td>inappropriate tone/connotation</td>
<td>misleading statement</td>
</tr>
<tr>
<td>The fact the college president invites black people to his party</td>
<td>black people</td>
<td>yes</td>
<td>This is completely biased against the individuals, as if the president should not act like this.</td>
<td>inappropriate tone/connotation</td>
<td>misleading statement</td>
</tr>
</tbody>
</table>
### Thematic analysis results.

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Keyword</th>
<th>Label</th>
<th>Comment</th>
<th>Code</th>
<th>Theme</th>
</tr>
</thead>
<tbody>
<tr>
<td>The fact the college president invites black people to his party.</td>
<td>black</td>
<td>people</td>
<td>Why should black people be invited to something shocking?</td>
<td>inappropriate tone/notation</td>
<td>misleading statement</td>
</tr>
<tr>
<td>How did the home and public lives of working class immigrant and African</td>
<td>white</td>
<td>women</td>
<td>It’s biased because it alludes to automatically only POC struggle and</td>
<td>inappropriate tone/notation</td>
<td>misleading statement</td>
</tr>
<tr>
<td>American women compare with wealthy middle and working class white women.</td>
<td></td>
<td></td>
<td>white people are well off</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Although slavery was not legal in the northern states free black residents</td>
<td>black</td>
<td>resident</td>
<td>Blacks were not given the same amount of rights, so I am going to say</td>
<td>misleading assumption</td>
<td>misleading statement</td>
</tr>
<tr>
<td>of the states were not given the same civil or political rights or treated with white people are well off</td>
<td></td>
<td></td>
<td>that it is biased</td>
<td></td>
<td></td>
</tr>
<tr>
<td>the same dignity and respect as white inhabitants</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discuss the role of the church as a place of local activism student involve-</td>
<td>whites</td>
<td>yes</td>
<td>It refers to supremacists, Ku Klux Klan as the opposition</td>
<td>negative association</td>
<td>negative language</td>
</tr>
<tr>
<td>ment and the roles of the opposition namely the Ku Klux Klan and Southern whites</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The fact they have black women computers hired to do this this important work</td>
<td>black</td>
<td>yes</td>
<td>There’s a negative feeling at the beginning of that sentence</td>
<td>negative tone</td>
<td>negative language</td>
</tr>
<tr>
<td>The franchise or right to vote was being extended to more white males as incomerelated eligibility requirements were being dropped by more states</td>
<td>white</td>
<td>yes</td>
<td>I feel like it is negatively biased on the grounds that they mention income. It feels as though they are negative about the fact that not every 'white male' would make a lot of money.</td>
<td>negative tone</td>
<td>negative language</td>
</tr>
<tr>
<td>The lines between refined white womanhood and degraded enslaved Black femaleness were no longer so clearly defined</td>
<td>white</td>
<td>yes</td>
<td>The use of “femaleness” against “womanhood” is somewhat dehumanising of black women.</td>
<td>inappropriate term</td>
<td>poor terminology</td>
</tr>
</tbody>
</table>
### Thematic analysis results.

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Keyword</th>
<th>Label</th>
<th>Comment</th>
<th>Code</th>
<th>Theme</th>
</tr>
</thead>
<tbody>
<tr>
<td>He believed that the United States governments both state and federal represented the interests of whites exclusively which made political participation by black Americans a waste of time at best and complicity in their own oppression at worst</td>
<td>whites</td>
<td>yes</td>
<td>Whites is a derogatory term similar to blacks despite them both being OK terms to use</td>
<td>inappropriate term</td>
<td>poor terminology</td>
</tr>
<tr>
<td>Do the illustrations depict nonwhites in subservient and passive roles or in leadership and action roles</td>
<td>non-whites</td>
<td>yes</td>
<td>Non-white is a derogatory term against people racialised as non-white by white people to incite whiteness as superior</td>
<td>inappropriate term</td>
<td>poor terminology</td>
</tr>
<tr>
<td>Do the illustrations depict nonwhites in subservient and passive roles or in leadership and action roles</td>
<td>non-whites</td>
<td>yes</td>
<td>The passive/active binary is not one but is, rather, relative, a series of related situations requiring perfect and total depiction of the racio types on one hand, but inviting scorn and mockery were they to be followed definitively.</td>
<td>inappropriate term</td>
<td>poor terminology</td>
</tr>
<tr>
<td>These documents chronicle a case in the wider wave of violence that targeted people of color during Reconstruction</td>
<td>people of color</td>
<td>yes</td>
<td>I always find the use of the word color in this manner to be a little distasteful in my opinion</td>
<td>inappropriate term</td>
<td>poor terminology</td>
</tr>
<tr>
<td>In the 1890s more than 100000 blacks were voting but by 1906 only 5000 managed to get through these barriers</td>
<td>blacks</td>
<td>yes</td>
<td>The phrasing ‘blacks’ seems to objectify and dehumanise black individuals. However, the discussion of breaking through barriers indicates that black people are able to overcome obstacles, which may be considered a positive thing.</td>
<td>inappropriate term</td>
<td>poor terminology</td>
</tr>
</tbody>
</table>
## Thematic analysis results.

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Keyword</th>
<th>Label</th>
<th>Comment</th>
<th>Code</th>
<th>Theme</th>
</tr>
</thead>
<tbody>
<tr>
<td>Texas shares with many other states especially with former Confederate</td>
<td>whites</td>
<td>yes</td>
<td>Referring to the individuals as their ethnic identification instead of as</td>
<td>inappropriate term</td>
<td>poor terminology</td>
</tr>
<tr>
<td>states a history of the systematic disenfranchisement of blacks Latinos</td>
<td></td>
<td></td>
<td>people. ex: blacks instead of black people. It makes the reference dehu-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>and poor whites</td>
<td></td>
<td></td>
<td>manizing</td>
<td></td>
<td></td>
</tr>
<tr>
<td>In the later 1960s the targets of Kings activism were less often the</td>
<td>blacks</td>
<td>yes</td>
<td>I am uncertain of the of the term 'blacks' which makes people appear</td>
<td>inappropriate term</td>
<td>poor terminology</td>
</tr>
<tr>
<td>legal and political obstacles to the exercise of civil rights by blacks</td>
<td></td>
<td></td>
<td>to be more like objects, than people.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>and more often the underlying poverty unemployment lack of education and</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>blocked avenues of economic opportunity confronting black Americans</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Describe the lives of free blacks and the laws that limited their</td>
<td>blacks</td>
<td>yes</td>
<td>Again the term “blacks” is an issue.</td>
<td>inappropriate term</td>
<td>poor terminology</td>
</tr>
<tr>
<td>freedom and economic opportunities</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tocqueville in a footnote gives two striking and even shocking</td>
<td>blacks</td>
<td>yes</td>
<td>There's an unsavoriness to the word “Blacks.” Like a more innocent way</td>
<td>inappropriate term</td>
<td>poor terminology</td>
</tr>
<tr>
<td>examples of majority tyranny at work the murder of antiwar journalists</td>
<td></td>
<td></td>
<td>of saying “Negros.” Even if this is used in a more historical context, it's</td>
<td></td>
<td></td>
</tr>
<tr>
<td>by a Baltimore mob and the intimidation of free blacks on Election Day in</td>
<td></td>
<td></td>
<td>not that hard to say “Black people.”</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Philadelphia</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nappy photos of black and brown people to help users be purposeful about</td>
<td>brown</td>
<td>yes</td>
<td>Nappy is a bad term</td>
<td>inappropriate term</td>
<td>poor terminology</td>
</tr>
<tr>
<td>representation in designs and presentations licensed under CC 0 public</td>
<td>people</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>domain dedication</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>They were often administered arbitrarily with more blacks required to</td>
<td>blacks</td>
<td>yes</td>
<td>By the use of black in and of itself</td>
<td>inappropriate term</td>
<td>poor terminology</td>
</tr>
<tr>
<td>take them than whites</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The Freedom Rides and attempts to integrate southern state universities</td>
<td>blacks</td>
<td>yes</td>
<td>Could have used black people instead of blacks.</td>
<td>inappropriate term</td>
<td>poor terminology</td>
</tr>
<tr>
<td>prompted him to deploy federal marshals in defense of blacks demanding</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>equal rights</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Thematic analysis results.

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Keyword</th>
<th>Label</th>
<th>Comment</th>
<th>Code</th>
<th>Theme</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Freedom Rides and attempts to integrate southern state universities prompted him to deploy federal marshals in defense of blacks demanding equal rights</td>
<td>blacks</td>
<td>yes</td>
<td>We reflect on history, we don’t have to use the exact terminology of history. Just say black people. “Blacks” is biased.</td>
<td>inappropriate term</td>
<td>poor terminology</td>
</tr>
<tr>
<td>According to the EDSITEmentreviewed website Women of the West Museum in the period from 17921844 the constitutions of Connecticut Delaware Kentucky Maryland New Jersey North Carolina Tennessee and Virginia excluded blacks from voting but expanded white male suffrage</td>
<td>blacks</td>
<td>yes</td>
<td>“Blacks” is demeaning in terms of the context of the sentence</td>
<td>inappropriate term</td>
<td>poor terminology</td>
</tr>
<tr>
<td>Others pointed to the example of Nat Turner a welltreated literate slave who instigated a rebellion in 1831 that resulted in the massacre of nearly sixty white men women and children before his capture and the deaths of almost two hundred blacks at the hands of white mobs</td>
<td>blacks</td>
<td>yes</td>
<td>White people are personalized while black people are phrased as just “blacks”. Also the emphasis on “well treated” is an attempt to dismiss the fact that Nat Turner was still an enslaved man who obviously despised his condition.</td>
<td>inappropriate term</td>
<td>poor terminology</td>
</tr>
<tr>
<td>Jim Crow Laws This term which came to be used to designate any law requiring racial segregation was borrowed from a racially stereotyped black character in a common nineteenthcentury songanddance act</td>
<td>black charac-ter</td>
<td></td>
<td>the fact it states ’stereotyped’ and ’common’</td>
<td>poor language</td>
<td>poor terminology</td>
</tr>
<tr>
<td>The decision to do so was a controversial one as many whites–no matter what their view of slavery–did not relish the thought of blacks bearing weapons fearing reprisals from armed former slaves resentful for their previous forced servitude</td>
<td>blacks</td>
<td>yes</td>
<td>dehumanising language</td>
<td>poor language</td>
<td>poor terminology</td>
</tr>
</tbody>
</table>
Thematic analysis results.

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Keyword</th>
<th>Label</th>
<th>Comment</th>
<th>Code</th>
<th>Theme</th>
</tr>
</thead>
<tbody>
<tr>
<td>This criticism of King was elaborated the following year by a fellow Baptist minister Joseph H Jackson president of the National Baptist Convention from 1953-1982 who delivered a speech counseling blacks to reject direct confrontation and stick to law and order.</td>
<td>blacks</td>
<td>yes</td>
<td>Feels like a poorly worded attempt at addressing a problem already caused by another person, it feels wrong.</td>
<td>poor language</td>
<td>poor terminology</td>
</tr>
<tr>
<td>Late in 1917 the US War Department created two allblack infantry divisions.</td>
<td>black infant</td>
<td>yes</td>
<td>They may have been feeding into the stereotypes that Black men are exceptionally strong and that their lives are worthless when making all Black infantry divisions</td>
<td>stereotypical language</td>
<td>poor terminology</td>
</tr>
<tr>
<td>When minority heroes do appear are they admired for the same qualities that have made white heroes famous or because what they have done have benefited white people.</td>
<td>white people</td>
<td>yes</td>
<td>This statement is biased because, although it is telling the uncomfortable truth but it is not filtered which means it is still be specifically mentioning certain people of ethnic group and showing, and comparing with other ethnic group.</td>
<td>unnecessary ethnic specification</td>
<td>unnecessary specification</td>
</tr>
<tr>
<td>About twelve years ago I hired a whaleboat and four black men and proceeded to Long Island after a load of round clams.</td>
<td>black men</td>
<td>yes</td>
<td>I don’t know why black men is mentioned specifically when the excerpt refers to them later as just men.</td>
<td>unnecessary ethnic specification</td>
<td>unnecessary specification</td>
</tr>
<tr>
<td>Why did Addams a white woman serve as the spokesperson for the NAACP?</td>
<td>white woman</td>
<td>yes</td>
<td>The sentence implies the white woman cannot serve because of her skin color.</td>
<td>unnecessary skin color specification</td>
<td>unnecessary specification</td>
</tr>
</tbody>
</table>
## Thematic analysis results.

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Keyword</th>
<th>Label</th>
<th>Comment</th>
<th>Code</th>
<th>Theme</th>
</tr>
</thead>
<tbody>
<tr>
<td>There is a similar intellectual maneuver in MerryGoAround instead of directly protesting the laws that say black people should ride in the back of trains a serious matter Hughes chooses to write about a merrygoround—which has no back</td>
<td>black</td>
<td>yes</td>
<td>I’m confused the need to add the colour of someone’s skin into this text.</td>
<td>unnecessary skin color specification</td>
<td>unnecessary specification</td>
</tr>
<tr>
<td>Many free blacks were able to become businessmen and leaders</td>
<td>blacks</td>
<td>yes</td>
<td>I can’t believe if this is true or not</td>
<td>no specific reason</td>
<td>yes - no reason</td>
</tr>
<tr>
<td>He believed that the United States governments both state and federal represented the interests of whites exclusively which made political participation by black Americans a waste of time at best and complicity in their own oppression at worst</td>
<td>whites</td>
<td>yes</td>
<td>Although the sentence is bias against white people the contents of it are definitely true and needed to be said</td>
<td>no specific reason</td>
<td>yes - no reason</td>
</tr>
<tr>
<td>This treaty was to bring peace between the whites and the Sioux who agreed to settle within the Black Hills reservation in the Dakota Territory</td>
<td>whites</td>
<td>yes</td>
<td>they only wanted peace</td>
<td>no specific reason</td>
<td>yes - no reason</td>
</tr>
<tr>
<td>No longer asserting that whites were devils but still skeptical of American institutions to secure the civil rights of black Americans Malcolm argued that the civil rights movement needed to be taken to a more receptive international forum such as the United Nations and World Court</td>
<td>whites</td>
<td>yes</td>
<td>Not exactly sure</td>
<td>no specific reason</td>
<td>yes - no reason</td>
</tr>
<tr>
<td>Malcolm X argued that America was too racist in its institutions and people to offer hope to blacks</td>
<td>blacks</td>
<td>no</td>
<td>About racism</td>
<td>context-exemption</td>
<td>context-exemption</td>
</tr>
</tbody>
</table>
## Thematic analysis results.

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Keyword</th>
<th>Label</th>
<th>Comment</th>
<th>Code</th>
<th>Theme</th>
</tr>
</thead>
<tbody>
<tr>
<td>This was Malcolm X's curt explanation of why he did not favor integration of blacks with whites in the United States</td>
<td>blacks</td>
<td>no</td>
<td>In Malcom X's explanation as to why he said that, I believe in his saying that America was too racist in its institutions and people to offer hope to blacks, what he meant was that 'It is not that we don't want to support blacks but it is already too late as the racism is seated deeply rooted in the America institutions and its people.</td>
<td>historical content</td>
<td>context-exemption</td>
</tr>
<tr>
<td>This bill required the President to appoint a provisional governor who would enroll all white males and then require these same men to take a so-called &quot;Ironclad oath&quot;—stating that they had never voluntarily borne arms against the United States or given “aid countenance counsel or encouragement” to the Confederacy—if they wanted to participate in electing delegates to a constitutional convention</td>
<td>white</td>
<td>male</td>
<td>historical context</td>
<td>historical content</td>
<td>context-exemption</td>
</tr>
<tr>
<td>Segregation means to keep black people and white people separate from one another</td>
<td>black</td>
<td>people</td>
<td>This isn’t bias because this definition in the context of American history is true.</td>
<td>historical content</td>
<td>context-exemption</td>
</tr>
<tr>
<td>Though Republicans made quick and huge political gains in the South as newly enfranchised black voters rushed to their support it was clear that the party and its ideals of peace through racial and sectional harmony on Republican terms remained unpopular with large segments of the population—particularly with those who were disenfranchised because they could not take the &quot;oath&quot; or otherwise prove their loyalty to the Union</td>
<td>black</td>
<td>voter</td>
<td>Because the passage discusses the history of an ethnic group, so colour needs to be made explicit, however it uses the term 'people' which humanises</td>
<td>historical content</td>
<td>context-exemption</td>
</tr>
</tbody>
</table>
### Thematic analysis results.

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Keyword</th>
<th>Label</th>
<th>Comment</th>
<th>Code</th>
<th>Theme</th>
</tr>
</thead>
<tbody>
<tr>
<td>Competing Voices of the Civil Rights Movement This lesson sequence presents the views of several important black leaders who shaped the debate over how to achieve freedom and equality in a nation that had long denied a portion of the American citizenry the full protection of their rights</td>
<td>black</td>
<td>no</td>
<td>The title is about black history</td>
<td>historical content</td>
<td>context-exemption</td>
</tr>
<tr>
<td>How are the Europeans views of Africans and the Africans views of whites represented in the novel</td>
<td>whites</td>
<td>no</td>
<td>the sentence should not make individuals referred to as “whites” feel negativity or exclusion. to deny the existence of whites in African literature would be to erase the truth of an african perspective and therefore people referred to as whites should not take offence or feel bias against them due to this</td>
<td>historical content</td>
<td>context-exemption</td>
</tr>
<tr>
<td>Profiles in Courage To Kill A Mockingbird and the Scottsboro Boys Trial–Students study select court transcripts and other primary source material from the second Scottsboro Boys Trial of 1933 a continuation of the first trial in which two young white women wrongfully accused nine AfricanAmerican youths of rape</td>
<td>white</td>
<td>no</td>
<td>This is factual and helps the reader understand what is going on.</td>
<td>factual statement</td>
<td>factual statement</td>
</tr>
<tr>
<td>They were often administered arbitrarily with more blacks required to take them than whites</td>
<td>whites</td>
<td>no</td>
<td>This is a fact, it cannot be bias.</td>
<td>factual statement</td>
<td>factual statement</td>
</tr>
<tr>
<td>It came in various sized from three to eight feet long and had previously been banned in the South by whites</td>
<td>whites</td>
<td>no</td>
<td>It had been banned previously, but it was now no longer banned, I dont think the statement is biased</td>
<td>factual statement</td>
<td>factual statement</td>
</tr>
</tbody>
</table>
Thematic analysis results.

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Keyword</th>
<th>Label</th>
<th>Comment</th>
<th>Code</th>
<th>Theme</th>
</tr>
</thead>
<tbody>
<tr>
<td>At the time there were separate schools for whites and blacks but the Court determined that this was not fair because having separate schools resulted in unequal opportunities for children to learn</td>
<td>blacks</td>
<td>no</td>
<td>not really, its explaining how things factually were during that time</td>
<td>factual statement</td>
<td>factual statement</td>
</tr>
<tr>
<td>Both blacks and whites are in this long line</td>
<td>blacks</td>
<td>no</td>
<td>It was bias at all but highlighted key facts</td>
<td>factual statement</td>
<td>factual statement</td>
</tr>
<tr>
<td>Competing Voices of the Civil Rights Movement This lesson sequence presents the views of several important black leaders who shaped the debate over how to achieve freedom and equality in a nation that had long denied a portion of the American citizenry the full protection of their rights</td>
<td>black leader</td>
<td>no</td>
<td>It seems to just be factual from a textbook.</td>
<td>factual statement</td>
<td>factual statement</td>
</tr>
<tr>
<td>What argument did some whites make for their move to Texas other then the geography of the region</td>
<td>whites</td>
<td>no</td>
<td>Once again no hidden agenda, just factual information.</td>
<td>factual statement</td>
<td>factual statement</td>
</tr>
<tr>
<td>About twelve years ago I hired a whaleboat and four black men and proceeded to LongIsland after a load of round clams</td>
<td>black men</td>
<td>no</td>
<td>It is a genuine story that happened and those happen to be the details</td>
<td>factual statement</td>
<td>factual statement</td>
</tr>
<tr>
<td>Martin Robison Delany was an editor author physician abolitionist black nationalist colonizationist and army officer</td>
<td>black nationalist</td>
<td>no</td>
<td>Factual statement</td>
<td>factual statement</td>
<td>factual statement</td>
</tr>
<tr>
<td>Martin Robison Delany was an editor author physician abolitionist black nationalist colonizationist and army officer</td>
<td>black nationalist</td>
<td>no</td>
<td>Seems a statement of fact rather than anything to get offended about include in a debate against slavery</td>
<td>factual statement</td>
<td>factual statement</td>
</tr>
<tr>
<td>the ways in which blacks and whites were to be separated</td>
<td>whites</td>
<td>no</td>
<td>Stating a fact.</td>
<td>factual statement</td>
<td>factual statement</td>
</tr>
<tr>
<td>Across the South Democrats controlled congressional apportionment based on total population although they had disenfranchised the black population</td>
<td>black population</td>
<td>no</td>
<td>No because ir is referring to the specific population that is Black</td>
<td>factual statement</td>
<td>factual statement</td>
</tr>
</tbody>
</table>
Thematic analysis results.

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Keyword</th>
<th>Label</th>
<th>Comment</th>
<th>Code</th>
<th>Theme</th>
</tr>
</thead>
<tbody>
<tr>
<td>No longer asserting that whites were devils but still skeptical of American institutions to secure the civil rights of black Americans Malcolm argued that the civil rights movement needed to be taken to a more receptive international forum such as the United Nations and World Court</td>
<td>whites</td>
<td>no</td>
<td>Considering the history of how the white people have treated the blacks, the sentence is not biased.</td>
<td>the target individual is privileged</td>
<td>group-exemption</td>
</tr>
<tr>
<td>In 2007 for the first time since the early nineteenth century Hispanics accounted for more than half of all births 502 while nonHispanic whites accounted for just 34</td>
<td>whites</td>
<td>no</td>
<td>White people have more systematic power so it isn’t biased</td>
<td>the target individual is privileged</td>
<td>group-exemption</td>
</tr>
<tr>
<td>Call attention to the story of the slave Randall that closes Browns first chapter pages 1619 in the electronic text and to episodes in which Brown himself offers resistance such as his snowball battle with a group of white boys Chapter 3 page 28 and his repeated attempts to escape Chapter 2 page 21 Chapter 7 pages 6568 and Chapter 10 pages 8993</td>
<td>white boy</td>
<td>no</td>
<td>It looks like a lesson plan. Did not seem overly positive or negative.</td>
<td>neutral sentence</td>
<td>neutral or not offensive language</td>
</tr>
<tr>
<td>Ruby was a black community organizer director in the federal Freedmens Bureau and leader of the Galveston Union League</td>
<td>black community</td>
<td>no</td>
<td>It is important for context as what community she is helping.</td>
<td>context description</td>
<td>neutral or not offensive language</td>
</tr>
<tr>
<td>For a visual image of the pursuit of civil rights by following principles of law and order have students access online a Charles Moore photograph of the registering of black voters in Mississippi</td>
<td>black voter</td>
<td>no</td>
<td>no, they are highlighting racial behavior</td>
<td>descriptive language</td>
<td>neutral or not offensive language</td>
</tr>
<tr>
<td>Nevertheless the fight to hold onto segregationist practices began to wear on some whites the question remained how best to address the concerns of local black citizens</td>
<td>whites</td>
<td>no</td>
<td>I’m not sure this is biased against either but more of an explanation of some happenings.</td>
<td>descriptive language</td>
<td>neutral or not offensive language</td>
</tr>
<tr>
<td>Sentence</td>
<td>Keyword</td>
<td>Label</td>
<td>Comment</td>
<td>Code</td>
<td>Theme</td>
</tr>
<tr>
<td>-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>---------</td>
<td>-------</td>
<td>--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>---------------------------</td>
<td>---------------------------</td>
</tr>
<tr>
<td>A census and reapportionment based on the number of white citizens was to be held every ten years</td>
<td>white</td>
<td>citizen</td>
<td>The author is simply describing the details of Texas’ government from 1866. It teaches about racism not encouraging.</td>
<td>descriptive language</td>
<td>neutral or not offensive language</td>
</tr>
<tr>
<td>Several hundred black men fought on the British side at the Battle of Great Bridge a minor engagement that took place on December 9 1775</td>
<td>black</td>
<td>men</td>
<td>descriptive of the soldiers</td>
<td>descriptive language</td>
<td>neutral or not offensive language</td>
</tr>
<tr>
<td>After completing their worksheets tell each group that they will take on the role of a committee in Congress examining the problems faced by blacks in the aftermath of the Civil War</td>
<td>blacks</td>
<td>no</td>
<td>It is not in any way bias, just addressing issues concerning a demography</td>
<td>fair description</td>
<td>neutral or not offensive language</td>
</tr>
<tr>
<td>How were black people and white people treated at this time</td>
<td>white</td>
<td>people</td>
<td>Black and white people are described in exactly the same way, save for the necessary racial distinction.</td>
<td>fair description</td>
<td>neutral or not offensive language</td>
</tr>
<tr>
<td>What minimum percentage of white male voters need to take the oath before those citizens could begin the process of reestablishing a new state government</td>
<td>white</td>
<td>male</td>
<td>I don’t think this is bias because it is asking for a percentage of male voters to reestablish new state government</td>
<td>fair question</td>
<td>neutral or not offensive language</td>
</tr>
<tr>
<td>When we let freedom ring when we let it ring from every village and every hamlet from every state and every city we will be able to speed up that day when all of Gods children black men and white men Jews and Gentiles Protestants and Catholics will be able to join hands and sing in the words of the old Negro spiritual Free at last free at last</td>
<td>white</td>
<td>men</td>
<td>The sentence clearly equates black men and white men as being alike.</td>
<td>fair statement</td>
<td>neutral or not offensive language</td>
</tr>
<tr>
<td>It has content and imagery relating to people of colour caucasian people and a picture of a Buddhist Monk</td>
<td>people of colour</td>
<td>no</td>
<td>seems objective and neutral</td>
<td>neutral sentence</td>
<td>neutral or not offensive language</td>
</tr>
</tbody>
</table>
## Thematic analysis results.

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Keyword</th>
<th>Label</th>
<th>Comment</th>
<th>Code</th>
<th>Theme</th>
</tr>
</thead>
<tbody>
<tr>
<td>Segregation means to keep black people and white people separate from one another</td>
<td>black</td>
<td>no</td>
<td>Seems neutral</td>
<td>neutral sentence</td>
<td>neutral or not offensive language</td>
</tr>
<tr>
<td>Formerly enslaved people in Galveston rejoiced in the streets after the announcement although in the years afterward many struggled to work through the changes against the resistance of whites</td>
<td>whites</td>
<td>no</td>
<td>The sentence alludes to the fact that enslaved people were still facing hardships by whites. Although I don’t think the excerpt is in favor or against “whites”</td>
<td>neutral sentence</td>
<td>neutral or not offensive language</td>
</tr>
<tr>
<td>The white clergymen complained that local black citizens were being directed and led in part by outsiders to engage in demonstrations that were unwise and untimely</td>
<td>black</td>
<td>no</td>
<td>I do not think this is a negative term - if anything it sounds inclusive using the word ‘citizen’</td>
<td>not negative</td>
<td>neutral or not offensive language</td>
</tr>
<tr>
<td>Many free African Americans particularly those in South Carolina Virginia and Louisiana were wealthy and well educated two facts that distinguished them from much of the white population both before and after the Civil War</td>
<td>white</td>
<td>no</td>
<td>It does not say anything negative about them but more information would help.</td>
<td>not negative</td>
<td>neutral or not offensive language</td>
</tr>
<tr>
<td>Tobias Gibson lamenting the mixing of Negro and white children in the same schoolroom</td>
<td>white</td>
<td>no</td>
<td>No, I think the term used to describe black children at the time had negative connotations, white children is still a term used and its never had negative connotations attached to it.</td>
<td>not negative</td>
<td>neutral or not offensive language</td>
</tr>
<tr>
<td>Why is Douglass specific about making friends with “little white boys”</td>
<td>white</td>
<td>no</td>
<td>No negative traits mentioned</td>
<td>not negative</td>
<td>neutral or not offensive language</td>
</tr>
<tr>
<td>One of these programs the Civilian Conservation Corps provided more than a quarter of a million young black men with jobs and was consequently another arena in which the black community waged the struggle for greater equality</td>
<td>black</td>
<td>no</td>
<td>there is nothing offensive here</td>
<td>not offensive</td>
<td>neutral or not offensive language</td>
</tr>
</tbody>
</table>
## Thematic analysis results.

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Keyword</th>
<th>Label</th>
<th>Comment</th>
<th>Code</th>
<th>Theme</th>
</tr>
</thead>
<tbody>
<tr>
<td>There was a general trend in the first half of the 19th century to extend the right to vote to more white males</td>
<td>white</td>
<td>male</td>
<td>Calling someone “white” is like the default in this day and age. Being white has become a default compared to being African American or Asian American. If you’re white in America, you’re just American. No bias.</td>
<td>not offensive</td>
<td>neutral or not offensive language</td>
</tr>
<tr>
<td>In this reflective essay Chief Dan George compares his native North American culture with that of white culture and expresses his view...</td>
<td>white</td>
<td>culture</td>
<td>This is just someone giving their view, they haven’t even given it, I don’t think this is biased.</td>
<td>opinion</td>
<td>neutral or not offensive language</td>
</tr>
<tr>
<td>In the years since Matthiessen’s important work and especially in the past several decades this characterization of the literary period has been challenged on several fronts especially for overstating the innovations of these few authors for the exclusion of women people of color and more popular authors from its account of the United States during a period of social and cultural upheaval and transition and for its acceptance of the myth of American exceptionalism or superiority</td>
<td>people of color</td>
<td>no</td>
<td>The sentence puts forward the idea that “people of color” have created many exceptional and notable works of literature that were overlooked in the past by Matthiesen, and is therefore uplifting them and not biased against them.</td>
<td>positive bias (not negative)</td>
<td>neutral or not offensive language</td>
</tr>
<tr>
<td>Commenorative sites continue to influence and shape our perception of history and the absence of public statues dedicated to female historic figures promulgates a narrow historical narrative that elevates the experiences and accomplishments of white men</td>
<td>white men</td>
<td>no</td>
<td>biased in favor of white men</td>
<td>positive bias (not negative)</td>
<td>neutral or not offensive language</td>
</tr>
</tbody>
</table>
Thematic analysis results.

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Keyword</th>
<th>Label</th>
<th>Comment</th>
<th>Code</th>
<th>Theme</th>
</tr>
</thead>
<tbody>
<tr>
<td>He was still dubious of the American political system but advised black Americans to 1 engage in smarter political voting and organization for example no longer voting for black leaders he viewed as shills for white interests and 2 fight for civil rights at the international level where he thought the nonwhite nations of the world would side with the oppressed black American minority and pressure the United States through the United Nations and World Court to protect their rights</td>
<td>non-white</td>
<td>nation</td>
<td>Though Poc is used more often than non-white, it doesn’t sound bias</td>
<td>proper tone</td>
<td>neutral or not offensive language</td>
</tr>
<tr>
<td>Another example of how Achebe foreshadows the alteration of indigenous society is the replacement by the white mans court of the clans customs with their own laws discussed in Chapter 20</td>
<td>white man</td>
<td>no</td>
<td>This is quoting the name of something and is not biased towards or against it.</td>
<td>quoting</td>
<td>neutral or not offensive language</td>
</tr>
<tr>
<td>Students should write a brief essay using as a starting point William Andrews assertion that the purpose of the slave narrative was to enlighten white readers about both the realities of slavery as an institution and the humanity of black people as individuals deserving of full human rights</td>
<td>black people</td>
<td>no</td>
<td>The question is not likely to cause black people to feel excluded or underrepresented and instead is is likely to make them feel perhaps vindicated and hopefully more understood. Educating people on the atrocities of the slave trade may be harrowing for black people but it is necessary for other ethnic groups</td>
<td>supportive language</td>
<td>neutral or not offensive language</td>
</tr>
<tr>
<td>I believe this because I have many black women in my college who are just as smart as white women and black or white men</td>
<td>black women</td>
<td>no</td>
<td>Suggesting equality in knowledge</td>
<td>supportive language</td>
<td>neutral or not offensive language</td>
</tr>
<tr>
<td>There were waiting rooms at the hospital to keep black and white people separate there were separate public restrooms water fountains and schools</td>
<td>white people</td>
<td>no</td>
<td>It may be bias against black people</td>
<td>no specific reason</td>
<td>no - no reason</td>
</tr>
</tbody>
</table>
## Thematic analysis results.

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Keyword</th>
<th>Label</th>
<th>Comment</th>
<th>Code</th>
<th>Theme</th>
</tr>
</thead>
<tbody>
<tr>
<td>The franchise or right to vote was being extended to more white males as incomerelated eligibility requirements were being dropped by more states</td>
<td>white</td>
<td>male</td>
<td>i cannot fathom how stating white men can vote is biased again white men</td>
<td>no specific reason</td>
<td>no - no reason</td>
</tr>
<tr>
<td>Appealing to the historic contribution of blacks to the development and prosperity of America Jackson counseled that less controversial and provocative means should be adopted in the struggle for civil rights</td>
<td>blacks</td>
<td>no</td>
<td>It is not at all bias</td>
<td>no specific reason</td>
<td>no - no reason</td>
</tr>
</tbody>
</table>