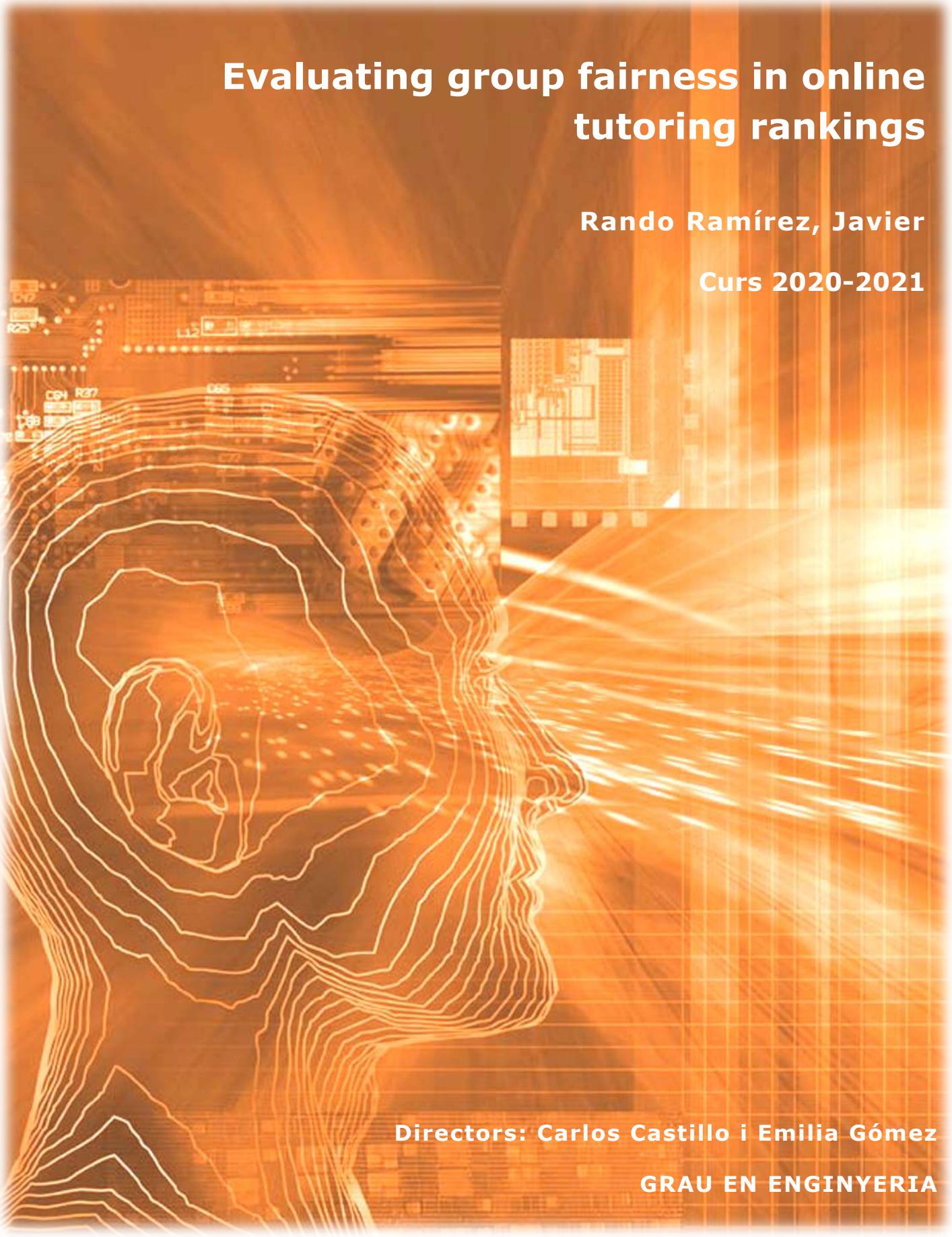

Evaluating group fairness in online tutoring rankings

Rando Ramírez, Javier

Curs 2020-2021

Directors: Carlos Castillo i Emilia Gómez

GRAU EN ENGINYERIA

upf. Universitat Pompeu Fabra Barcelona
Escola d'Enginyeria

Treball de Fi de Grau

Bachelor Degree Thesis

Universitat Pompeu Fabra

Evaluating group fairness in online tutoring rankings

Javier Rando Ramírez

Supervisors: Carlos Castillo and Emilia Gómez

June 2021

A mis padres y mi hermano por no poner límites a mi imaginación.

Acknowledgments

I would like to sincerely thank everyone who made this work possible. On the first place, Carlos Castillo for awakening my interest in the application of Data Science for social good and give me the opportunity to learn from him during all these years. Thanks to Emilia Gómez for supporting me in this work and always demand a little more from me. Thanks to María Sánchez del Río for helping me to express my work mathematically, for putting up with engineers' misuse of mathematics, and for many more things that would not fit in this paper. Thanks to my colleagues, especially Sara Estévez, and Eduard Vergés, for all these years of suffering, learning and, above all, laughter. Thanks to Javi Castillo and June Monge for being them. And finally, thanks to my family for bringing me here.

Abstract

This project addresss the evaluation of *group fairness* in commercial search engines that may lead to discrimination against certain groups of individuals. More precisely, we measure fairness with respect to gender and nationality in a set of platforms where users can search for second language teachers completely online. These websites must rank available teachers for visitors, and disparate exposure may lead to uneven economic benefit. We evaluate if teachers belonging to protected groups are fairly represented in the first positions of the rankings, and we conduct a statistical analysis of price depending on nationality and gender across languages and platforms. Our results put forward that although most lists are fair for all groups, there are some worrisome exceptions. Finally, we found out that women and people from high-income countries charge significantly higher fees for their classes.

Keywords: algorithmic fairness, ranking bias, data mining, information retrieval, machine learning, language tutoring.

Contents

Contents iv

List of Figures vi

List of Tables vii

1 Introduction and theoretical background 1
- 1.1 Research Questions . . . 3
- 1.2 Report Structure . . . 4
- 1.3 Related work . . . 5
- 1.4 Ranking problem formulation . . . 11

2 Methodology 12
- 2.1 Data acquisition and preparation . . . 12
  - 2.1.1 Italki . . . 13
  - 2.1.2 Preply . . . 19
  - 2.1.3 Verbling . . . 22
- 2.2 Selecting languages . . . 26
- 2.3 Including income level for teachers . . . 26
- 2.4 Inferring gender . . . 28
  - 2.4.1 Building ground-truth dataset . . . 29
  - 2.4.2 Comparing different libraries . . . 29
  - 2.4.3 Handling missing values . . . 31
  - 2.4.4 Gender inference limitations . . . 31

3 Results 33
- 3.1 FA*IR Analysis . . . 33
  - 3.1.1 Gender . . . 34

3.1.2 Nationality 36
3.1.3 Discussion 38
3.2 Statistical price analysis 41
3.2.1 Gender 41
3.2.2 Nationality 43
3.2.3 Nationality using Big Mac Index 46
3.2.4 Gender and nationality 48
3.2.5 Discussion 50
3.3 Modeling ranking 51
3.3.1 Discussion 53

4 Conclusion **57**

5 Final remarks **60**
5.1 Limitations and future work 60
5.2 Ethical considerations 61

Bibliography **63**

A Languages available per platform **71**

B Italki languages grouping **74**

C Nationalities in Preply **75**

D Kolmogorov-Smirnov Test **77**

E Reproducibility: code and data **79**

List of Figures

1.1 Research stages 4

2.1 Sample teacher JSON retrieved from Italki API. 18
2.2 Preply filter from which all available languages were retrieved. Allows scroll down to access more elements in the list. 19

List of Tables

2.1 Online language tutoring platforms detailed comparison. 13
2.2 Teacher attributes considered for Italki and their type and description. 16
2.3 Mapping from the JSON representation of teachers to the fields specified in Table 2.2. 17
2.4 Teacher attributes considered for Preply and their type and description. 22
2.5 Teacher attributes considered for Preply and their type and description. 25
2.6 GDP per capita (US$) in 2019 for a set of countries and regions. 28
2.7 Performance metrics for gender inference tools 30
2.8 Performance metrics for gender inference using custom decision boundaries 31

3.1 Sample FA*IR analysis 34
3.2 FA*IR analysis for top-10 ranking and women as the protected attribute 36
3.3 FA*IR analysis for top-100 ranking and women as the protected attribute 37
3.4 FA*IR analysis for top-10 ranking and low-income countries as the protected attribute 39
3.5 FA*IR analysis for top-100 ranking and low-income countries as the protected attribute 40
3.6 Significance test for price using women as the protected attribute in Italki 42
3.7 Significance test for price using women as the protected attribute in Preply 43

3.8 Significance test for price using women as the protected attribute in Verbling . . . 43
3.9 Significance test for price using low-income countries as the protected attribute in Italki . . . 44
3.10 Significance test for price using low-income countries as the protected attribute in Preply . . . 45
3.11 Significance test for price using low-income countries as the protected attribute in Verbling . . . 45
3.12 Significance test for Big Macs using low-income countries as the protected attribute in Italki . . . 47
3.13 Significance test for Big Macs using low-income countries as the protected attribute in Preply . . . 47
3.14 Significance test for Big Macs using low-income countries as the protected attribute in Verbling . . . 48
3.15 Average price with respect to gender and nationality on Italki . 49
3.16 Average price with respect to gender and nationality on Preply 49
3.17 Average price with respect to gender and nationality on Verbling 50
3.18 Attributes considered for position regression across platforms . . 53
3.19 Attributes available for each plaftorm . . . 54
3.20 Correlation matrix for attributes in Table 3.18 considering all languages for Italki. . . . 54
3.21 Correlation matrix for attributes in Table 3.18 considering all languages for Preply. . . . 55
3.22 Correlation matrix for attributes in Table 3.18 considering all languages for Verbling. . . . 55
3.23 Regression metrics for position on Italki. . . . 55
3.24 Regression metrics for position on Preply. . . . 56
3.25 Regression metrics for position on Verbling. . . . 56

A.1 Number of teachers for languages considered for the analysis in each platform . . . 72
A.2 Number of teachers for languages not considered for the analysis in each platform . . . 73

B.1 Grouped languages in Italki to match other platforms. . . . 74

C.1 Number of unique teachers per nationality in Preply. Nationalities are represented using their corresponding ISO codes. . . . 76

Chapter 1

Introduction and theoretical background

During the last decades, the use of the Internet has increased worldwide. In Europe, in 1990 only 0.06% of the population had access to the Internet. In 2019, this percentage was already 85.42% [1]. The Internet has changed the way we live and interact. It enables, among others, instant communication worldwide, shopping in almost any store from home and accessing any information from a mobile device.

This transformation has resulted, among others, in the generation of huge amounts of unstructured data. We usually refer to it as Big Data [42]. Its retrieval, transformation, processing and usage is widely studied in the fields of Data Mining [36] and Information Retrieval [5]. One of the main challenges is to design techniques that obtain meaningful information from it [42].

[1] Data retrieved from The World Bank dataset "Individuals using the Internet". Available at: https://data.worldbank.org/indicator/it.net.user.zs?locations=EU

At the same time, the increasing computational power and data available have boosted the development of Artificial Intelligence (AI). This technology has the potential to learn using data without specific instructions or guidance [61]. It is becoming extremely useful because it can outperform humans in specialized tasks such as playing chess [65] or medical diagnosis [50]. Also, it is widely used by institutions to automate tasks and extract information from huge amounts of data [24]. However, some risks arise from AI and it may harm individuals in the short and long-term future [52, 10, 60].

AI systems are starting to support sensitive decisions like selecting students for college, providing insurance, or sentencing in the courts [52]. If these systems behave differently for distinct groups of people, they may become a threat for democracy and equality. Ensuring equal opportunities for everyone is one of the main goals of modern societies [30]. Hence, measuring fairness in AI systems dealing with personal data is extremely important. Given the opacity and complexity of algorithms and the expensive data collection, it is usually costly and time-consuming to find whether a system is biased against a group of people. Recently, scandals regarding people classification by artificial intelligence have made clear how important it is to evaluate these tools before deploying them in the real world. For instance, journalists from ProPublica discovered how an algorithm used in the USA to determine how likely was someone to commit a crime was biased against black people [4]. Also, Reuters reported that Amazon had been using a recruiting tool that discriminated against women [23].

This project focuses on evaluating fairness in people search engines; i.e. machine learning systems that create the best ranking from a collection of people to answer a query. More precisely, we measure diversity in online

language tutoring platforms where people post their profiles as language teachers. Then, visitors can hire language lessons completely online. Since teaching is done remotely, people from different nationalities, locations, ethnicities, and genders can offer the same service. Each platform must rank the candidates to provide a list in which potential clients navigate and choose their teacher.

The likelihood that an entry is visited in a ranking is highly influenced by its position. According to the *Google Organic CTR Report* provided by Advanced Web Ranking[2], more than 71% of organic searches on Google resulted on a click on the first page results and the first 5 entries got 67.60% of the visits. Thus, enforcing fair ranking tools for people is really important to provide equal opportunities for everyone. In the context of online language tutoring, disparate exposure may lead to uneven economic benefits between groups.

1.1 Research Questions

This work assesses whether the ordering in three online language tutoring platforms (Italki, Preply, and Verbling) is fair. We evaluate if women and people from low-income countries are treated similarly to men and teachers from high-income countries. We choose these protected features because poverty is strongly linked to discrimination and poor health conditions [57] and women are still underrepresented in the labour market and earn less money than men [19].

[2]More information at: https://www.advancedwebranking.com/blog/google-organic-ctr/

Our analysis tries to put forward potential bias in rankings, and it also explores whether a significant price difference exists between groups. Although the price is chosen directly by teachers, it might be lower for certain groups of users due to previous prejudices. Finally, we try to reproduce underlying black-box ranking algorithms using public data to obtain further insights about features importance.

1.2 Report Structure

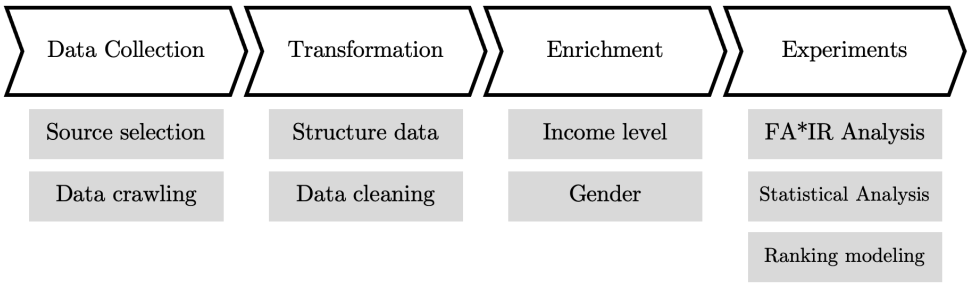

Figure 1.1: Research stages

This report is divided into 5 chapters and an appendix. At first, we present the motivation, research questions and related work. Then, we detail the research stages (see Figure 1.1) followed to obtain experimental data and generate results. At the end, we include final remarks and conclusion. The report structure is the following:

- **Introduction and theoretical background**. Current section. It presents the motivation for the project, research questions to be assessed, and a summary of related work on the topic.
- **Methodology**. A detailed explanation of data extraction, transfor-

mation, and enrichment.

- **Results**. Evaluation of ranking fairness, statistical price differences, and algorithms approximation.
- **Conclusion**. Sums up the analysis and insights obtained throughout the process.
- **Final remarks**. Puts forward the work constraints, future research ideas, and an ethical discussion on our methodology.

1.3 Related work

Ranking in Information Retrieval

Search engines [21] help users find the most relevant content to meet their information needs. More formally, given a query q, search engines try to sort *documents* d_i within a collection $D = \{d_1, d_2, ..., d_n\}$ by relevance. First, each document is assigned a $Score(q, d_i)$ which is computed using a function of query and document that reflects relevance. Finally, an ordered list $R = [d_{(1)}, d_{(2)}, ..., d_{(n)}]$ is constructed such that $Score(q, d_{(1)}) > Score(q, d_{(2)}) > ... > Score(q, d_{(n)})$.

To compute $Score(q, d_i)$, the so-called ranking algorithms are used [43, 21, 5]. Some of these algorithms are based purely on deterministic relevance metrics between query and content. For instance, when dealing with text documents and queries, one of the most popular similarity metrics is the term frequency-inverse document frequency (TF-IDF) [55, 68, 62]. In more complex contexts such as the web, additional information can be extracted from statistics such as PageRank to provide more accurate results [54].

However, more sophisticated approaches to ranking implement machine learning models as $Score(q, d_i)$ for better results. They allow to consider more features and generalize for unseen scenarios. This field is called *learning to rank* [43]. It is the most common technique in real-world ranking problems and search engines started to use it in the early 2000s when Altavista presented a gradient-boosting algorithm [58]. Ranking can be performed using supervised and unsupervised machine learning techniques. In supervised algorithms, there is a rating for each of the results, usually given by a domain expert, and this information is used to train a ranking model [28]. On the other hand, unsupervised machine learning algorithms may learn from scratch using content and/or user interaction metrics to determine relevance [74].

Fairness and bias in decision-making

Humans face daily relevant decisions which may affect other individuals. Unfortunately, studies show that unrelated external factors or stereotypes can influence even egalitarian people without their knowledge [48]. For example, fatigue, hunger or time of the day can influence judges at court [22, 16]. Also, racial and socio-economic status disparities can lead to misdiagnosis in medical decisions [34, 67, 64, 47].

In this work, we will distinguish between discrimination, (un)fairness and bias. *Discrimination* is a legal concept and refers to situations in which individuals are treated differently based on their membership to a certain group [2]. The right to nondiscrimination is a essential part in the European Union normative framework [33]. There exists *fairness* in a process when there is no discrimination against individuals. However, *bias* is a technical concept that describes a deviation from a state which is known to be true

[25]. Bias can lead to discrimination against certain groups [33].

Another relevant term used throughout this report is *protected feature*. It is a legal term that defines an attribute based on which it is not legal to discriminate against. In the scope of this project, we use gender and country of origin as protected characteristics according to Article 14 of the European Convention on Human Rights[3][51].

Algorithmic fairness

So-called algorithmic decision systems (ADS) raised as an useful tool to support human decision-making [13, 66, 41]. This technology uses machine learning algorithms that, usually, learn from historical data to accurately predict unseen situations. Their objective is to maximize precision and their decisions are based purely on the facts [66, 61]. However, evidence puts forward that they can reproduce human biases in their predictions [6]. For example, Correctional Offender Management Profiles for Alternative Sanctions (COMPAS) is a well-known tool that was used in the United States of America to determine criminals' likelihood to re-offend. This likelihood was then used by a judge, along with further information, to decide whether people should be remanded in custody. Journalists from ProPublica showed that its predictions were biased against African Americans [4, 39].

Thus, it is really important to identify potential biases in ADS before deploying them. *Algorithmic fairness* tries to mitigate potential biases in machine learning problems that may lead to discrimination while preserving

[3]Article 14 of the European Convention of Human rights considers "sex, race, colour, language, religion, political or other opinion, national or social origin, association with a national minority, property, birth or other status" as protected attributes.

accuracy [20, 66]. Fairness in automated processes is usually measured as *individual fairness* or *group fairness*. Individual fairness [26] ensures that all individuals with similar features obtain a similar output. On the other hand, group fairness [56] defines a set of protected features (e.g. black people, women, etc.) and tries that, in general, the set of samples containing these protected features are treated similarly to the remaining population. Although they might look similar, there is an open discussion on whether both individual and group fairness can be fulfilled at the same time in practice [7].

A common group fairness metric based on predictions is *demographic parity* (see Equation 1.1). The idea underlying this method is that the outcome is independent of the protected attribute. More formally, the expected outcome ($\hat{y}$) for both groups (G) must be equal to ensure fairness [66].

$$\mathbb{E}[\hat{y} = 1|G = a] = \mathbb{E}[\hat{y} = 1|G = b] \tag{1.1}$$

Discrimination against protected groups and potential solutions have been explored in different domains [46] such as image classification [40, 49], natural language processing [9, 44, 15], hiring tools [59], justice [39] or people ranking [3, 29].

Fairness in Ranking

This project is focused on evaluating group fairness in machine learning algorithms used to rank people in information retrieval scenarios. It has been shown that it is possible to find discrimination in rankings against certain

individuals due to arbitrary reasons such as gender [17]. Even when these attributes are not directly considered as input to machine learning models, implicit information in different features [14] like language corpora [11] can result in biased predictions. Zehlike et al. [73] present a vast overview of state-of-the-art fairness considerations when building ranking algorithms.

Once a ranking is provided to a visitor, the likelihood that an entry is visited decreases as its position increases. The *Google Organic CTR Report* provided by Advanced Web Ranking[4] puts forward than more than 71% of organic searches on Google resulted on a click on first page results. What is more, the first 5 entries got 67.60% of the visits. Hence, disparate exposure for different groups may lead to uneven outcomes.

Previous works have proposed several metrics for fairness in ranked outputs. Yang and Stoyanovich [70] compare distributions of protected and non-protected items in different top-k lists and apply a discounted average. Zehlike et al. [72] defined and solved the *Fair Top-k Ranking problem*. Their solution is the first algorithm that can identify and mitigate bias in ranked lists using statistical tests. In Section 3.1, we use their tool (FA*IR) to assess group fairness in our dataset. Previous publications have already assessed fairness in ranked outputs using this framework [71, 31].

Some researchers have started to put forward unfair ranking systems and potential solutions. For instance, Geyik et al. [32] propose a framework that ensures fairness while preserving business metrics evaluated on LinkedIn Talent Search. Researchers have also studied discrimination when ranking

[4]More information at: https://www.advancedwebranking.com/blog/google-organic-ctr/

jobs and workers in Google and TaskRabbit [3]. Elbassuoni et al. [29] evaluate discrimination in online job marketplaces and present subgroup fairness by considering any combination of protected attributes. Finally, Galindo [31], a fellow student at Pompeu Fabra University, explored the potential biases of commercial search engines. Her work is used as inspiration for our analysis.

Online language tutoring

This work focuses on online language learning platforms since they provide a rich context in which people from different locations, ethnicities and genders offer the same service. Also, their impact is relevant since online tools are becoming a popular mean to learn a second language [8, 35]. In fact, according to a study[5], the online language tutoring market is expected to grow at a CAGR of 18.7% from 2020 and reach a $21.2 billion value by 2027.

Assessing fairness on these platforms is important because, as presented before, the number of visits an entry gets is highly dependent on its position in the list. In online tutoring websites, the teachers' goal is to make money. Their exposure to visitors will determine how many potential students they have. Therefore, unfair orderings based on protected features could result in potential income inequality for people belonging to protected groups.

[5]Market research study conducted by Meticulous Researcn. More information at: https://www.meticulousresearch.com/product/online-language-learning-market-5025

1.4 Ranking problem formulation

This section presents a formal definition of the ranking problem and the nomenclature we will use throughout the report. The ranking problem tries to sort a set of *documents* $D = \{d_1, d_2, ..., d_n\}$ according to their relevance to a query q. For this, each document is assigned a $Score(q, d_i)$ computed as a function of document and query that measures relevance. Items are then returned ordered by descending score.

In the context for this analysis, each language, $lang$, is a set D^{lang} where items, d_i^{lang}, represent all teachers offering lessons for that language. Furthermore, D^{lang} is obtained from the union of two disjoint sets containing non-protected (NP) and protected (P) teachers respectively. Thus, $D^{lang} = D_{NP}^{lang} \bigsqcup D_P^{lang}$.

Our setup considers that the query q is the language from which items will be retrieved. Finally, $Score(q, d_i)$ is an unknown black-box function that returns a sorted list of elements $R = [d_{(1)}^{lang}, d_{(2)}^{lang}, ..., d_{(n)}^{lang}]$ such that $Score(q, d_{(1)}^{lang}) > Score(q, d_{(2)}^{lang}) > ... > Score(q, d_{(n)}^{lang})$.

Chapter 2

Methodology

2.1 Data acquisition and preparation

In this chapter, the source selection, data extraction, and its transformation process used throughout the project are presented. The first step was selecting the sources for experimental data. In order to retrieve the information automatically, sources needed to provide public access without paid memberships and lax measures against crawling. Furthermore, they must have a large number of teachers to obtain reliable conclusions.

Through web search, we identified the following platforms offering online language lessons: Italki, Verbling, Preply, VerbalPlanet, Lingoda and Rype. However, only Italki[1], Preply[2] and Verbling[3] allowed automated information retrieval, and they all had more than 40 active languages. VerbalPlanet had mainly unstructured data, and Lingoda and Rype required paid membership. Detailed comparison is presented in Table 2.1

[1]https://italki.com
[2]https://preply.com
[3]https://verbling.com

Table 2.1: Online language tutoring platforms detailed comparison.

	Italki	**Preply**	**Verbling**	**VerbalPlanet**	**Lingoda**	**Rype**
Launched	2007	2012	2011	2006	2013	2016
Country of origin	China	Ukraine	USA	UK	Germany	USA
Languages	42	41	40	36	N/A	N/A
Paid membership	No	No	No	No	Yes	Yes
Structured data	Yes	Yes	Yes	No	N/A	N/A

Then, a crawling strategy was designed for each of these platforms. The goal is to obtain as much information as possible to conduct an in-depth analysis of rankings in all of them. Since each of the websites has a completely different structure, three different processes to scrap the information are considered.

2.1.1 Italki

Extraction

After performing a network traffic analysis, we discovered that information was retrieved by the interface from a public API. Therefore, this API was used directly to obtain the data without handling interaction with the browser.

Firstly, available languages in the platform were retrieved. API responses contained a unique code for all languages in Italki. Another endpoint (`https://translate.italki.com/i18n/en_us.json`) provided a mapping between language codes and full language names.

Language codes are required to query the API for information. Following the requests performed by the website, a request template was built to retrieve the listings from Italki:

- Method: `POST`
- URL: `https://api.italki.com/api/v2/teachers`
- Headers: {`"Content-Type": "application/json"`}
- Body: {`"teach_language":{"language": [LANGUAGE_CODE]}, "page_size": [NUMBER_TO_RETRIEVE], "user_timezone": "Europe/Madrid"`}

An important remark on this extraction is that more than 100 items cannot be retrieved at the same time. To overcome this limitation for languages with more than 100 teachers, paging must be considered in the requests. Therefore, it was required to iterate through pages containing 20 teachers until the whole list was retrieved. The body of the request was modified accordingly:

- Body: {`"teach_language":{"language": [LANGUAGE_CODE]}, "page_size": "20", "page": [PAGE_ITERATOR], "user_timezone": "Europe/Madrid"`}.

  `[PAGE_ITERATOR]` is a number increasing from 1 up to $\lceil \frac{\text{num\_teachers}}{20} \rceil$.

Also, it is important to wait for some seconds between calls to ensure responsible usage of resources and to avoid IP blocking. Since there are no specifications in the `robots.txt` file, the following rules were considered:

- 3 seconds between consecutive languages.
- Random uniform number in the range $[0, 2]$ between pages of the same language.

The whole JSON files for each language obtained during this process

were stored locally for further transformation. This way, data is only crawled once.

Transformation

After collecting data for all teachers in the platform as presented in the previous section, JSON files returned by the API were structured into tables. A sample teacher JSON entry is shown in Figure 2.1. To obtain the desired information specified in Table 2.2, teachers' JSON was transformed using the mapping presented in Table 2.3.

Table 2.2: Teacher attributes considered for Italki and their type and description.

Field	Type	Description
position	Integer	Position in ranking
language	String	Language crawled
retrieval_date	Datetime	Date of data retrieval
user_id	String	Unique ID for the user
user_name	String	User full name (may include free text)
avatar_file_name	String	Avatar filename
video_picture	String	URL of the video miniature
is_pro	Boolean	Whether user is PRO
origin_country	String	Nationality
teaches	Dictionary	Languages taught and level
also_speaks	Dictionary	Languages spoken by teacher
in_platform_since	Datetime	Day in which teacher registered
rating	Float	Average rating
number_sessions	Integer	Total number of classes in Italki
price	Float	Price per session
price_time	String	Class duration
price_currency	String	Currency for the price

Table 2.3: Mapping from the JSON representation of teachers to the fields specified in Table 2.2.

Field	**Value in JSON**
position	Position in list (not in JSON)
language	Language of crawling (not in JSON)
retrieval_date	Date of crawling (not in JSON)
user_id	`['user_info']['user_id']`
user_name	`['user_info']['nickname']`
avatar_file_name	`['user_info']['avatar_file_name']`
video_picture	`['teacher_info']['qiniu_video_pic_url']`
is_pro	`['user_info']['is_pro']`
origin_country	`['user_info']['origin_country_id']`
teaches	`['teacher_info']['teach_language']`
also_speaks	`['teacher_info']['also_speak']`
in_platform_since	`['teacher_info']['first_valid_time']`
rating	`['teacher_info']['overall_rating']`
number_sessions	`['teacher_info']['session_count']`
price	`['course_info']['min_price']`
price_time	'hour' (not in JSON)
price_currency	USD (not in JSON)

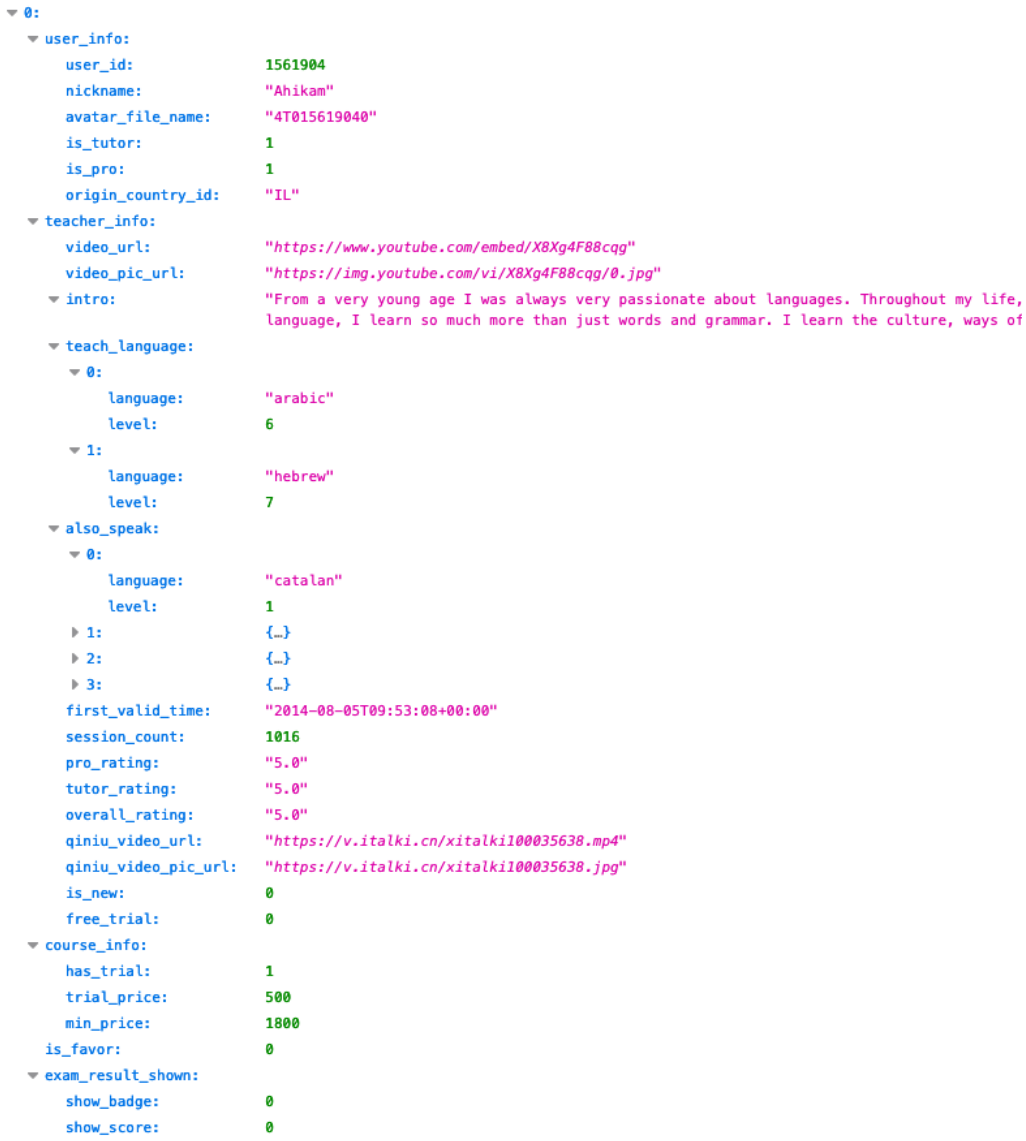

Figure 2.1: Sample teacher JSON retrieved from Italki API.

2.1.2 Preply

Extraction

In this platform, no public API was available so a scraper that simulates browser interaction was built to obtain the information. The package used for this purpose is Selenium [4]. It launches a browser, Google Chrome in this work, and gives developers full control over interaction and retrieval.

As in Italki, the first step was collecting all available languages on the platform. The easiest way was visiting the page for a specific language and retrieving all other languages from the filter (see Figure 2.2).

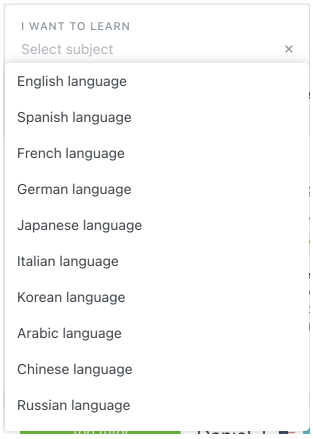

Figure 2.2: Preply filter from which all available languages were retrieved. Allows scroll down to access more elements in the list.

Once the list was retrieved, `language` was removed from the string and the listing URLs to access teacher information were constructed following this structure: `https://preply.com/en/skype/[LANGUAGE]-tutors`.

[4]Details can be found at: https://selenium-python.readthedocs.io/index.html

Obtaining teachers information was really time-consuming so it was useful to split the task into two different steps that eased the process:

1. **Retrieving ranking information**. In this first step, the whole ranking was scrapped storing teacher name, URL, and position. This way, rankings for all languages are stored at the same time to avoid the effect of unconsidered variables such as date, time, etc.
2. **Asynchronous teachers details retrieval**. Then, using all stored URLs for teachers, they were asynchronously visited. The whole HTML content for each teacher was stored locally for faster processing.

Scrapping challenges

While retrieving information from Preply, several challenges were encountered.

- **Avoid IP blocking**. Preply is really strict when it comes to limiting requests to the server. This is the main reason why the process was split into different subtasks for efficient retrieval. Requests to the server should be time spaced. After trying several combinations and experiencing blocks, it was found out that the following values worked properly. Uniform random number of seconds in the range [120, 180] before loading a new language. Between 7 and 10 after accessing a new page in the listing for a given language. And between 1 and 2 after retrieving a teacher HTML page. Sleeping times must be randomized to get closer to real user behavior.
- **Handling errors**. After many requests, the platform may stop returning desired content for a while. This was common when retrieving

HTML content for teachers since a lot of requests are required. In this case, the whole process was stopped for a random uniform time between 20 and 30 seconds and retried the request.

- **Maximum number of teachers**. Given request limitations, it was not possible to retrieve the whole content for languages containing many teachers. 600 is the maximum number of teachers to consider.

Transformation

The transformation process was more complicated than for Italki because it required structuring information from a raw HTML file. The fields considered for this platform are displayed in Table 2.4. Each of the fields required accessing different elements in the HTML. The main problem with this approach is that it is highly dependent on the website version available. In fact, while preparing the scrapper, an update in the platform completely changed the elements.

Table 2.4: Teacher attributes considered for Preply and their type and description.

Field	Type	Description
position	Integer	Position in ranking
language	String	Language crawled
retrieval_date	Datetime	Date of data retrieval
user_name	String	User full name (may include free text)
avatar_url	String	Avatar URL
url	String	Teacher URL in the platform
is_featured	Boolean	Whether user is featured in ranking
nationality_full	String	Country of origin (full)
nationality	String	Country of origin (code)
teaches	List	All languages taught by the teacher
subjects	List	List of subjects taught for the language
speaks	Dictionary	Languages spoken by teacher and the level
avg_rating	Float	Average rating
num_ratings	Integer	Number of ratings received
lessons	Integer	Total number of classes in Preply
price	Float	Price per session
price_currency	String	Currency for the price

2.1.3 Verbling

Extraction

Crawl information from Verbling followed a very similar approach to Preply. No public API was available so it was required to simulate a browser and extract information directly from the platform. First, all available languages

were obtained by getting elements in the language filter. Then, each language ranking was visited, teachers' metadata and URLs were stored and finally, HTML content for each teacher was downloaded asynchronously.

Each language URL was constructed following the pattern used by the platform. In this case: `https://verbling.com/find-teachers?language=[LANGUAGE]`

Scrapping challenges

The main challenges faced when obtaining information were related to their tools to avoid scrapping.

- **Infinite scroll**. On each language page, the ranking is presented using an infinite scroll. New teachers are loaded when scrolling down. Also, it was required to ensure when the list finishes. A long loading time might be incorrectly interpreted as the end of the list since no new teachers showed up. For this, if no new teachers were found, the process was stopped for a random uniform number between 2 and 4 seconds. Then, tried to scroll down and slept again between 2 and 4 seconds. If no information was loaded, then it is assumed that the list came to an end.
- **Avoid IP blocking**. Like Preply, they are really strict controlling requests to their server. Therefore, work was again split into independent tasks, and sleeping times were designed to get closer to human behavior. Waiting times were introduced when updating the teacher list using scroll, between requests to get HTML pages for teachers, and before starting a new language crawl. To load new teachers, the

execution stopped between 2 and 4 seconds. Between HTML requests, between 1 and 2 seconds were elapsed and before scrapping a new language, between 5 and 10.

- **Design change**. They also implement a strategy to deter scrappers which consisted of changing the HTML tags after consecutive requests. User-friendly tags such as `"div.FindTeacherFilter-teaches"` were replaced by different identifiers like `"div.sc-jzJRlG"`. This broke the process because elements were no longer found. To solve this issue, waiting times are really important. However, the new identifiers are not random but constant every time so the script also supports the alternative tags on the website.

Transformation

It is required to structure HTML files into tables. The main shortcoming of this technique, as explained for Preply, is the high dependence on website version and elements. Even small changes in the platform can make obsolete the code. The values considered for Verbling are presented in Table 2.5

Table 2.5: Teacher attributes considered for Preply and their type and description.

Field	Type	Description
position	Integer	Position in ranking
language	String	Language crawled
retrieval_date	Datetime	Date of data retrieval
first_name	String	Teacher first name
last_name	String	Teacher last name
avatar_url	String	Avatar URL
url	String	Teacher URL in the platform
is_featured	Boolean	Whether user is featured in ranking
location	String	Location of the user (independent from nationality)
nationality	String	Country of origin (code)
dialect	String	Dialect of the language taught (optional)
class_details	List	List of target groups, specialties...
teaching_levels	List	List of levels for teaching (1 to 6)
speaks	Dictionary	Languages spoken by teacher and the level
avg_rating	Float	Average rating
num_ratings	Integer	Number of ratings received
avg_lessons_per_student	Float	Average number of lesson per student
lessons	Integer	Total number of classes in Preply
students	Integer	Total number of unique students in Preply
price	Float	Price per session
price_detail	Dictionary	Prices for sessions packs
price_currency	String	Currency for the price

2.2 Selecting languages

Although we are interested in measuring fairness for every language available in each platform, we focus the analysis on a subset to ensure that conclusions can be compared across websites and that the population is representative enough. We consider only languages that (1) are available on all platforms at the same time, and (2) have at least 10 teachers on every source.

The number of teachers can be interpreted as a measure of popularity for a certain language. Since languages with the highest number of teachers are those that will have a larger impact due to higher demand, we defined a smaller subset called TOP-LANGUAGES containing those in which disparity would be most harmful. TOP-LANGUAGES are those which have more than 100 teachers on average across platforms (see Table A.1).

There are 40 common languages among the websites. However, many of them have less than 10 teachers in some platforms. Hence, only 22 languages are considered in the scope of this work. English is the most popular language, with more than twice as many teachers as the second most popular. Spanish, French and Chinese are next in popularity. 12 languages are selected as TOP-LANGUAGES. Tables in Annex A display the total number of teachers available for each language and source, the average, and the selection we performed for the analysis.

2.3 Including income level for teachers

One important protected attribute when it comes to fairness is race as it was pointed out, among others, by the Charter of Fundamental Rights of the European Union [30]. For this reason, analyzing whether rankings or

attributes may discriminate certain groups based on their nationality was considered one of the main goals of this project.

It is required to define a protected group for which we want to ensure parity. We use GDP per capita to divide countries. The following income levels defined by the World Bank based on GDP per capita (2019) [5] are used:

- Low income: GDP per capita $\leq$ 810.1 USD
- Low middle income: 810.1 USD $<$ GDP per capita $\leq$ 2174.4 USD
- Middle income: 2,174.4 USD $<$ GDP per capita $\leq$ 9,013.7 USD
- Upper middle income: 9,013.7 USD $<$ GDP per capita $\leq$ 44,612.5 USD
- High income: 44,612.5 USD $<$ GDP per capita.

As a reference, some countries have been included in Table 2.6 with their income level. Protected nationalities are those within the following groups: low income, low middle income, and middle income. Thus, nationalities were aggregated as follows:

- Low income (protected): GDP per capita $\leq$ 9,013.7 USD
- High income: 9,013.7 USD $<$ GDP per capita.

Since the world's GDP per capita is 11,433.2 USD, all protected nationalities are below this value. The entire work will consider this binary classification of countries regarding teachers' origin countries.

[5]Information retrieved from: https://data.worldbank.org/indicator/NY.GDP.PCAP.CD

Table 2.6: GDP per capita (US$) in 2019 for a set of countries and regions.

Country	GDP per capita (USD)	Income group
Sudan	441.5	Low income
Afghanistan	507.1	Low income
Senegal	1,446.8	Low middle income
India	2,099.6	Low middle income
Nigeria	2,229.9	Middle income
Cuba	8,821.8	Middle income
Colombia	6,428.7	Middle income
China	10,216.6	Upper middle income
World	11,433.2	Upper middle income
Uruguay	16,190.1	Upper middle income
Spain	29,600.4	Upper middle income
European Union	34,913.2	Upper middle income
United Kingdom	42,330.1	Upper middle income
United States of America	65,297.5	High income
Switzerland	84,096.4	High income

2.4 Inferring gender

Gender is also considered a protected attribute according to the Charter of Fundamental Rights of the European Union [30]. Therefore, this work tries to find potential negative impact on women. However, none of the platforms provide this information explicitly in the public data. This section presents the process to find the best way to infer teachers' gender from available information. Two main approaches were considered:

- **Name**. Induce gender from teacher name and surname.
- **Profile picture**. Obtain gender using each teacher image.

To decide which was the best way to tackle this problem, three different libraries were compared using a manually tagged dataset as ground-truth. This setup allows to measure accuracy for this specific problem.

2.4.1 Building ground-truth dataset

Preply was used to build the ground-truth because it is the one with the largest number of nationalities. In fact, teachers from 129 different nationalities are found in Preply (118 in Italki and 80 in Verbling). Since gender inference tools can be very sensitive to the country of origin [37], we perform stratified sampling to fairly consider as many nationalities as possible. All nationalities and their corresponding number of teachers are represented in Annex C.

As manually tagging all teachers was unfeasible, the following heuristics were used to decide which nationalities to include and how many teachers to sample:

- Nationalities containing at least 4 teachers.
- At most 5 teachers per nationality.

After setting the nationalities and the number of teachers, users were randomly selected from the existing data. This resulted in a ground-truth of 359 teachers that were manually labeled using their profile pictures. If the decision was not clear, their information page was visited and gender was extracted from their description context. Precision is considered to be 100%. Overall, there were 238 females and 121 males.

2.4.2 Comparing different libraries

The next step was to compare how different libraries performed on the ground-truth dataset. We consider three different inference tools:

Table 2.7: Performance metrics for each of the gender inference tools. The Chosen library has been highlighted.

		Genderize	Gender API	DeepFace
Precision	Males	90%	88%	35%
	Females	98%	99%	68%
Recall	Males	68%	92%	60%
	Females	76%	86%	44%
F1-Score	Males	77%	90%	44%
	Females	86%	92%	53%

- Genderize[6]. API that allows inferring gender for a name. Nationality can also be included to narrow down the search.
- Gender API[7]. A very similar tool to Genderize. However, it enables the user to input a name with non-relevant words or surnames and it can extract the actual first name after a cleaning process.
- DeepFace [63]. This library infers gender from images. In this case, profile pictures are used as input.

Metrics obtained for each of them are presented in Table 2.7. DeepFace was discarded given the bad results obtained. Between Genderize and Gender API, the second one was selected because it has a higher recall. Nevertheless, metrics for both groups are unbalanced and males have higher recall. We tackle this issue by shifting the decision cutoff based on the probability of being male. After trying several values (see Table 2.8), people were labeled as male if they had a probability of being men greater or equal to 70%. Otherwise, they were tagged as women.

[6]Details about the library: https://genderize.io
[7]Details about the library: https://gender-api.com

Table 2.8: Performance metrics for each of the different decision boundaries. Probabilities of being men are represented. Probability for women can be computed as $1 - prob(men)$. Chosen decision boundary has been highlighted.

		0.5	0.6	0.7	0.8
Precision	Males	88%	91%	95%	96%
	Females	99%	98%	97%	95%
Recall	Males	92%	90%	88%	85%
	Females	86%	88%	90%	91%
F1-Score	Males	90%	90%	91%	90%
	Females	92%	92%	93%	93%

2.4.3 Handling missing values

Finally, using Gender API and the modified cutoff value gender was inferred on the whole dataset. Nonetheless, there were entries for which gender could not be inferred. These were the missing values for each platform:

- Italki - 259 missing values out of 7684 entries (3.37%)
- Preply - 57 missing values out of 6590 entries (0.7%)
- Verbling - 62 missing values out of 2116 entries (2.93%)

Missing values were introduced manually since it was feasible and it ensured high precision. Hence, gender is available for all entries in the dataset.

2.4.4 Gender inference limitations

Although a performance evaluation of the inference was performed, this process may induce biases in our analysis and results. To reduce its impact, we made sure that precision and recall for output classes are balanced, and we built the test set using stratified sampling on nationalities since accuracy might not be constant across them [38].

Also, inferring gender reduces the possible values to male and female. However, gender is not binary [45] and non-binary minorities cannot be taken into account.

Chapter 3

Results

3.1 FA*IR Analysis

This first analysis focuses on whether rankings are fair for protected groups using the FA*IR library [72]. In other words, we try to measure potential biases in the function $Score(q, d_i)$. FA*IR calculates if a proportion p of protected candidates ($d_i \in D_P^{lang}$) is fairly represented in the top-k ranking using statistical tests [71, 31]. We can define a function FA*IR(R, k, p) that takes as inputs ranking R, first k values to consider, and proportion p; and returns $True$ if a proportion p of protected teachers is fairly represented in top-k items of R, or $False$ otherwise.

In order to check if a ranking is biased, we find the highest proportion $p \in \{0.1, 0.2, ..., 1\}$ of protected teachers which is fairly represented in the first k positions in ranking R according to FA*IR. This value is set as p_{max} (see Equation 3.1). A sample output from our process is shown in Table 3.1. Then, we compare p_{max} with the total percentage of teachers, $\%prot$, belonging to the protected group in D^{lang} (see Equation 3.2). This comparison allows to assess potential bias in the ranking. If p_{max} is greater than $\%prot$, it means

Table 3.1: Sample output from a FA*IR analysis designed for this work. It is tested whether each proportion p of protected teachers is fairly represented in the list. FA*IR(R, k, p) returns $True$ (T) or $False$ (F). Highest true proportion is selected as p_{max}, in this case 0.6 (60%).

p	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Is fair?	T	T	T	T	T	T	F	F	F	F

that protected teachers are overrepresented in the top-k. On the other hand, if p_{max} is smaller than $\%prot$, it denotes underrepresentation of protected individuals. This analysis was performed for gender (see Section 3.1.1) and nationality (see Section 3.1.2) on each platform.

$$\begin{gathered} A = \{p \in \{0.1, 0.2, ..., 1\} : \text{FA*IR}(R, k, p) = True\} \\ p_{max} = \max_{p \in A} p \end{gathered} \tag{3.1}$$

$$\%prot = \frac{|D_P^{lang}|}{|D^{lang}|} \tag{3.2}$$

To establish a consistent framework for bias assessment, we will consider that a top-k list is biased whenever $p_{max} <= \%prot - 0.1$. This heuristic is chosen because there is a 0.1 step between consecutive proportions p. Since FA*IR already implements statistical tests to handle uncertainty, a 0.1 difference is significant enough.

3.1.1 Gender

In these experiments, the goal is to check whether women are fairly represented in the first positions of the rankings for each language. To get a broader picture of the rankings, we measured fairness in top-10 (see Table 3.2) and top-100 (see Table 3.3). In lists containing less than 100

teachers, top-10 and the whole set were considered. Finally, the highest fair proportions are compared against the overall women percentage in the listings.

Results in Table 3.2 show that, in general, women are fairly represented in top-10 rankings across languages and platforms. Nevertheless, there are some languages for which we spot biased results (4 on Italki and 1 on Preply). The most remarkable is German in Italki. The first 10 positions would only be fair if women accounted for 30% of the dataset. However, 56% of German teachers are women in Italki. We can also find three additional unfair rankings in Italki: Portuguese, Dutch and Persian. Preply only has one unfair ranking for Portuguese and Verbling is fair for all considered languages.

Table 3.3 collects the results for the first 100 positions. This broader picture allows to check whether the disparity is consistent in the whole dataset, or only arises for the first positions. At first sight, we notice that more rankings were classified as unfair (6 on Italki and 2 on Verbling). The majority are again found on Italki. German and Dutch results did not improve for the first 100 positions so women are definitely less likely to appear in first results than men. On the other hand, Portuguese no longer shows a significant difference. It is also worth mentioning that unfairness appeared in French, the third most popular language across platforms. Verbling has now two unfair rankings (Indonesian and Thai). Since they both have less than 100 teachers, these results let us conclude that the whole ranking for both languages is unfair for women. Preply has no biased rankings this time.

Table 3.2: Maximum p for which the FA*IR test is $True$ in top-10 ranking. Considers $\alpha = 0.1$ and females as the protected group. The original proportion of females (%prot) for the whole set of teachers is used as the fair baseline. Languages for which $p_{max} <= \%prot - 0.1$ have been highlighted. Separator after TOP-LANGUAGES.

Language	**Italki**		**Preply**		**Verbling**	
	p_{max}	%prot	p_{max}	%prot	p_{max}	%prot
English	0.4	0.45	0.7	0.63	0.7	0.51
Spanish	0.7	0.58	0.8	0.69	0.8	0.61
French	0.7	0.51	0.8	0.59	0.8	0.58
Chinese	0.9	0.68	0.9	0.66	0.8	0.71
Italian	0.6	0.65	0.9	0.67	0.7	0.69
Russian	0.7	0.78	0.9	0.81	0.8	0.85
Portuguese	0.4	0.51	0.5	0.59	0.8	0.61
Arabic	0.5	0.39	0.5	0.46	0.6	0.41
Japanese	0.8	0.65	0.9	0.63	0.8	0.74
German	0.3	0.56	0.8	0.62	0.7	0.49
Korean	0.6	0.61	0.6	0.56	0.8	0.60
Turkish	0.4	0.46	0.8	0.57	0.7	0.38
Polish	0.7	0.64	0.9	0.66	0.8	0.64
Hindi	0.5	0.41	0.7	0.62	0.8	0.40
Dutch	0.2	0.42	0.7	0.54	0.7	0.56
Greek	0.6	0.64	0.8	0.74	0.9	0.83
Indonesian	0.9	0.68	0.6	0.52	0.6	0.67
Persian	0.4	0.56	0.5	0.42	0.7	0.70
Thai	0.7	0.66	0.9	0.66	0.8	0.80
Vietnamese	0.6	0.57	0.8	0.69	0.7	0.70
Hebrew	0.4	0.42	0.6	0.47	0.8	0.60
Romanian	0.8	0.69	0.8	0.56	0.8	0.77

3.1.2 Nationality

After measuring fairness based on gender, we repeated the same procedure using income level for teachers. Now, low-income teachers are the protected group considered for analysis. Again, we considered top-10 (see Table 3.4) and top-100 (see Table 3.5). Unlike gender, some languages have no low-income teachers. Thus, results are not available for them.

Table 3.3: Maximum p for which the FA*IR test is $True$ in top-k ranking for $k = min(100, num\_teachers)$. Considers $\alpha = 0.1$ and females as the protected group. The original proportion of females (%prot) for the whole set of teachers is used as the fair baseline. Languages for which $p_{max} <= \%prot - 0.1$ have been highlighted. Separator after TOP-LANGUAGES.

Language	**Italki**		**Preply**		**Verbling**	
	p_{max}	%prot	p_{max}	%prot	p_{max}	%prot
English	0.5	0.45	0.7	0.63	0.7	0.51
Spanish	0.6	0.58	0.7	0.69	0.6	0.61
French	0.4	0.51	0.7	0.59	0.6	0.58
Chinese	0.6	0.68	0.7	0.66	0.7	0.71
Italian	0.5	0.65	0.8	0.67	0.7	0.69
Russian	0.8	0.78	0.9	0.81	0.8	0.85
Portuguese	0.5	0.51	0.5	0.59	0.6	0.61
Arabic	0.2	0.39	0.4	0.46	0.4	0.41
Japanese	0.7	0.65	0.7	0.63	0.8	0.74
German	0.3	0.56	0.6	0.62	0.5	0.49
Korean	0.5	0.61	0.5	0.56	0.6	0.60
Turkish	0.5	0.46	0.6	0.57	0.5	0.38
Polish	0.6	0.64	0.7	0.66	0.8	0.64
Hindi	0.4	0.41	0.6	0.62	0.6	0.40
Dutch	0.3	0.42	0.6	0.54	0.6	0.56
Greek	0.7	0.64	0.7	0.74	0.9	0.83
Indonesian	0.7	0.68	0.6	0.52	0.5	0.67
Persian	0.5	0.56	0.5	0.42	0.7	0.70
Thai	0.6	0.66	0.7	0.66	0.7	0.80
Vietnamese	0.6	0.57	0.8	0.69	0.8	0.70
Hebrew	0.5	0.42	0.5	0.47	0.8	0.60
Romanian	0.7	0.69	0.7	0.56	0.8	0.77

Table 3.4 display results for the first 10 teachers in each platform. Overall, people from low-income countries are fairly represented across languages and platforms. However, the first results for English in Italki are not representing protected individuals fairly. This is very relevant since it is the most popular language on all platforms.

Again, to validate whether these differences in a larger set, Table 3.5

presents results for the first 100 teachers. Results are again fair overall but parity for English in Italki did not improve. Thus, we can put forward that people from low-income countries are less likely to appear in the first results for English tutoring.

3.1.3 Discussion

Altogether, results from the analysis show that platforms display a fair ordering of teachers for most languages when considering gender and nationality as the protected attributes. Nevertheless, teachers belonging to these groups are less likely to appear in the first positions in some scenarios. This lets us think that fairness is not being assured by these websites at all times. *Italki* is the platform containing most of the unfair listings for women, and the only one offering a biased ranking against teachers from low-income countries. This last disparity is especially important because it shows up in English. It is the most popular language across platforms (most registered teachers) and, therefore, the amount of users that can be harmed is higher.

Table 3.4: Maximum p for which the FA*IR test is $True$ in top-10 ranking. Considers $\alpha = 0.1$ and low-income countries as the protected group. The original proportion of people from low-income countries (%prot) for the whole set of teachers is used as the fair baseline. Languages for which $p_{max} <= \%prot - 0.1$ (10%) have been highlighted. Languages for which there are no teachers with protected nationality have been discarded (-). Separator after TOP-LANGUAGES.

Language	**Italki**		**Preply**		**Verbling**	
	p_{max}	%prot	p_{max}	%prot	p_{max}	%prot
English	0.20	0.34	0.70	0.34	0.20	0.13
Spanish	0.80	0.32	0.40	0.23	0.30	0.25
French	0.20	0.17	0.30	0.23	0.30	0.09
Chinese	0.20	0.02	0.20	0.02	-	-
Italian	0.20	0.04	0.20	0.05	-	-
Russian	0.40	0.34	0.30	0.35	0.40	0.18
Portuguese	0.90	0.79	0.90	0.69	0.80	0.79
Arabic	0.90	0.90	0.90	0.92	0.90	0.95
Japanese	0.20	0.02	0.40	0.03	0.20	0.02
German	0.20	0.08	0.50	0.20	-	-
Korean	0.20	0.02	-	-	-	-
Turkish	0.20	0.11	0.50	0.10	-	-
Polish	0.20	0.08	0.20	0.15	-	-
Hindi	0.90	0.99	0.90	0.98	0.90	0.93
Dutch	0.20	0.05	0.30	0.04	-	-
Greek	0.20	0.03	0.30	0.06	-	-
Indonesian	0.90	0.94	0.90	0.94	1.00	1.00
Persian	0.90	0.91	0.50	0.25	0.90	0.83
Thai	0.90	0.97	0.90	0.98	0.90	0.85
Vietnamese	0.90	0.94	0.90	0.86	0.90	0.85
Hebrew	0.20	0.03	0.20	0.05	-	-
Romanian	0.30	0.17	0.60	0.38	0.30	0.15

Table 3.5: Maximum p for which the FA*IR test is $True$ in top-k ranking for $k = min(100, num\_teachers)$. Considers $\alpha = 0.1$ and low-income countries as the protected group. The original proportion of people from low-income countries (%prot) for the whole set of teachers is used as the fair baseline. Languages for which $p_{max} <= \%prot - 0.1$ (10%) have been highlighted. Languages for which there are no teachers with protected nationality have been discarded (-).

Language	**Italki**		**Preply**		**Verbling**	
	p_{max}	%prot	p_{max}	%prot	p_{max}	%prot
English	0.20	0.34	0.30	0.34	0.20	0.13
Spanish	0.40	0.32	0.20	0.23	0.30	0.25
French	0.10	0.17	0.20	0.23	0.10	0.09
Chinese (Mandarin)	0.00	0.02	0.00	0.02	-	-
Italian	0.00	0.04	0.20	0.05	-	-
Russian	0.40	0.34	0.30	0.35	0.40	0.18
Portuguese	0.90	0.79	0.90	0.69	0.80	0.79
Arabic	0.90	0.90	0.90	0.92	0.90	0.95
Japanese	0.00	0.02	0.00	0.03	0.00	0.02
German	0.10	0.08	0.50	0.20	-	-
Korean	0.00	0.02	-	-	-	-
Turkish	0.10	0.11	0.50	0.10	-	-
Polish	0.10	0.08	0.20	0.15	-	-
Hindi	0.90	0.99	0.90	0.98	0.90	0.93
Dutch	0.10	0.05	0.00	0.04	-	-
Greek	0.00	0.03	0.10	0.06	-	-
Indonesian	0.90	0.94	0.90	0.94	1.00	1.00
Persian (Farsi)	0.90	0.91	0.50	0.25	0.90	0.83
Thai	0.90	0.97	0.90	0.98	0.90	0.85
Vietnamese	0.90	0.94	0.90	0.86	0.90	0.85
Hebrew	0.00	0.03	0.20	0.05	-	-
Romanian	0.20	0.17	0.60	0.38	0.30	0.15

3.2 Statistical price analysis

In this section, we present a statistical analysis of price. Even when rankings are fair, protected groups can be discriminated in online platforms. Since people use these websites to get a job, we are interested in knowing whether users belonging to protected groups are likely to earn less money for the same job. Although price is fixed by candidates themselves, existing biases in society may lead them to offer lower prices. The experiment is conducted for top-40 entries, and the whole ranking. In those languages containing less than 40 teachers, only one test was performed using the whole list.

Analysis is performed using the Kolmogorov-Smirnov Test, which is a non-parametric test used to determine whether two independent empirical samples come from the same distribution. More details about how Kolmogorov-Smirnov Test is implemented can be found in Appendix D. We will consider price for protected teachers and non-protected teachers as two independent samples in our analysis. $X = (x_1, x_2, ..., x_{n_1})$ represents price for teachers belonging to D_{NP}^{lang}, and $Y = (y_1, y_2, ..., y_{n_2})$ for protected teachers within D_{P}^{lang}.

Statistical analysis of price using Kolmogorov-Smirnov Test will be presented considering gender and country of origin as protected features hereafter.

3.2.1 Gender

We try to determine whether women are likely to earn less money than men for the same service. Each platform was studied independently. Results for Italki (see Table 3.6), Preply (see Table 3.7) and Verbling (see Table

Table 3.6: Significance test to check whether **price** (USD) for men and women come from the same distribution in Italki. Results with p-value≤0.1 and at least 10 samples per group are represented. Hypothesis test conducted using Kolmogorov-Smirnov test. The disadvantaged group has been highlighted. Languages sorted by decreasing average number of teachers and separator after TOP-LANGUAGES.

Language	**Top-K**	**Males**		**Females**		**p-value**
		mean	median	mean	median	
Spanish	400	11.97	10.00	12.74	11.00	0.087
French	400	15.07	13.25	17.98	17.00	0.000
Russian	400	9.93	9.00	11.62	10.00	0.009
Portuguese	352	11.15	9.50	11.71	10.00	0.007
Arabic	220	11.22	10.00	13.71	10.00	0.069
Turkish	40	8.80	8.00	11.25	11.00	0.021
	158	9.19	8.25	10.76	10.00	0.013
Hindi	95	7.25	6.40	8.68	8.00	0.098
Dutch	78	14.56	13.45	18.66	16.30	0.030
Persian	40	8.05	7.60	11.44	10.00	0.058
	102	8.13	7.00	9.62	9.00	0.051

3.8) show a clear pattern. In most languages, prices come from the same distribution and when it is not the case, men are always the affected group with the lowest fares.

It is worth mentioning that, as we saw in Section 3.1, Italki is again the platform in which more disparities are found while Verbling is the fairest.

Table 3.7: Significance test to check whether **price** (EUR) for men and women come from the same distribution in Preply. Results with p-value≤0.1 and at least 10 samples per group are represented. Hypothesis test conducted using Kolmogorov-Smirnov test. The disadvantaged group has been highlighted. Languages sorted by decreasing average number of teachers and separator after TOP-LANGUAGES.

Language	Top-K	Males		Females		p-value
		mean	median	mean	median	
French	887	15.73	14.00	19.39	17.00	0.000
Portuguese	354	10.67	8.00	12.43	13.00	0.004
Arabic	420	9.86	8.00	11.08	10.00	0.022
Japanese	352	15.54	13.00	15.73	14.00	0.015
Polish	122	13.69	12.00	15.85	15.00	0.075
Persian	24	8.79	8.00	18.30	17.00	0.001
Hebrew	40	24.58	24.00	31.38	31.00	0.045
	59	24.86	24.00	31.97	30.00	0.054

Table 3.8: Significance test to check whether **price** (USD) for men and women come from the same distribution in Verbling. Results with p-value≤0.1 and at least 10 samples per group are represented. Hypothesis test conducted using Kolmogorov-Smirnov test. The disadvantaged group has been highlighted. Languages sorted by decreasing average number of teachers and separator after TOP-LANGUAGES.

Language	Top-K	Males		Females		p-value
		mean	median	mean	median	
Spanish	516	15.84	15.00	17.43	17.00	0.001
German	67	29.31	29.00	35.89	33.00	0.069

3.2.2 Nationality

Since this work assesses disparities for gender and nationality, the same experiment was reproduced considering teachers from low-income countries as the protected group. Again, results for Italki (see Table 3.9), Preply (see Table 3.10) and Verbling (see Table 3.11) are consistent across languages and platforms. However, the protected group is now the one with the lowest prices. Verbling is, once more, the platform showing greater parity and Italki repeats as the most unbalanced (14 languages with significant differences).

Table 3.9: Significance test to check whether **price** (USD) in Italki for high and low-income nationalities come from the same distribution. Results with p-value≤0.1 and at least 10 samples per group are represented. Hypothesis test conducted using Kolmogorov-Smirnov test. The disadvantaged group has been highlighted. Languages sorted by decreasing average number of teachers and separator after TOP-LANGUAGES.

Language	**Top-K**	**High-income**		**Low-income**		**p-value**
		mean	median	mean	median	
English	400	17.9	16.0	11.1	10.0	0.000
Spanish	400	13.3	12.0	10.4	9.9	0.000
French	40	17.8	19.0	8.9	9.0	0.007
	400	17.8	16.0	10.6	10.0	0.000
Italian	400	14.8	13.0	11.7	11.0	0.053
Russian	400	11.7	10.0	10.3	10.0	0.094
Portuguese	352	15.7	14.0	10.3	9.8	0.000
Arabic	220	24.6	18.0	10.9	10.0	0.000
Japanese	400	15.7	14.5	10.0	10.0	0.011
German	40	25.6	20.0	10.1	9.0	0.035
	400	20.8	18.0	13.1	12.0	0.000
Indonesian	65	11.9	11.8	8.1	8.0	0.018
Persian (Farsi)	40	14.5	14.5	9.1	8.0	0.038
Thai	40	18.0	18.0	10.2	10.0	0.050
	79	16.5	16.5	10.6	10.0	0.043
Vietnamese	89	16.8	14.0	9.2	9.0	0.004
Hebrew	65	18.7	16.0	10.0	10.0	0.088

The three most popular languages across platforms (English, Spanish and French) show big price differences depending on teachers' nationality. For instance, low-income English teachers in Verbling earn $15.15 while high-income instructors charge $22.89, in average. At the same time, within the first 40 entries for French in Preply, protected teachers get 15.50€ per lesson, and lecturers from high-income countries obtain 25.44€.

Although teachers from low-income countries are clearly earning less money for their lessons, the next subsection tries to assess this difference taking into account varying living costs across countries in the world.

Table 3.10: Significance test to check whether **price** (EUR) in Preply for high and low-income nationalities come from the same distribution. Results with p-value≤0.1 and at least 10 samples per group are represented. Hypothesis test conducted using Kolmogorov-Smirnov test. The disadvantaged group has been highlighted. Languages sorted by decreasing average number of teachers and separator after TOP-LANGUAGES.

Language	**Top-K**	**High-income**		**Low-income**		**p-value**
		mean	median	mean	median	
English	40	14.63	13.00	9.30	9.50	0.000
	591	17.56	16.00	12.72	12.00	0.000
Spanish	575	14.77	13.00	11.79	11.00	0.000
French	40	25.44	24.00	15.50	15.00	0.002
	887	19.57	17.00	12.78	12.00	0.000
Portuguese	40	16.63	17.00	14.38	13.00	0.027
	354	14.33	13.00	10.61	8.00	0.000
Arabic	420	13.71	13.00	10.20	8.00	0.030
German	462	23.19	21.00	16.09	15.00	0.000
Polish	122	15.82	15.00	11.00	11.00	0.003

Table 3.11: Significance test to check whether **price** (USD) in Verbling for high and low-income nationalities come from the same distribution. Results with p-value≤0.1 and at least 10 samples per group are represented. Hypothesis test conducted using Kolmogorov-Smirnov test. The disadvantaged group has been highlighted. Languages sorted by decreasing average number of teachers and separator after TOP-LANGUAGES.

Language	**Top-K**	**High-income**		**Low-income**		**p-value**
		mean	median	mean	median	
English	424	22.89	21.00	15.15	14.50	0.000
Spanish	516	17.47	17.00	14.95	15.00	0.002
French	157	25.95	25.00	14.85	12.00	0.000
Portuguese	89	19.58	20.00	14.97	15.00	0.000

3.2.3 Nationality using Big Mac Index

In Section 3.2.2, a statistical significant difference was spotted for price when comparing teachers from low and high-income countries. Nevertheless, since living costs vary greatly between these countries, we establish a common framework for comparison. The Big Mac Index established by The Economist in 1986[1] provides the price for a Big Mac hamburger in countries around the world. Since it is a service offered by the same company around the world, it has been proved to be a good metric to assess purchasing power parity [53].

For this analysis, we convert lessons price to the number of Big Macs that teachers could buy in their origin countries with the money earned for a lesson. The main limitation of this method is that we must assume that teachers live in their origin countries because current location is not provided by platforms. Then, statistical analysis is conducted to check whether the number of Big Macs for teachers from high and low-income countries come from the same distribution. Results differ from those obtained using directly price for Italki (see Table 3.12), Preply (see Table 3.13) and Verbling (see Table 3.14). Although teachers from low-income countries are, in general, still earning less in comparison with those from high-income countries, there are fewer unfair listings. Besides, in some languages, low-income teachers are benefited. For instance, English low-income teachers earn more money on every platform. This insight is really relevant since English is the most popular language and many teachers are affected.

[1]More details can be found at: https://www.economist.com/big-mac-index

Table 3.12: Significance test to check whether **price** (adjusted to Big Mac price per nationality) in Italki for high and low-income nationalities come from the same distribution. Results with p-value≤0.1 and at least 10 samples per group are represented. Hypothesis test conducted using Kolmogorov-Smirnov test. The disadvantaged group has been highlighted. Languages sorted by decreasing average number of teachers and separator after TOP-LANGUAGES.

Language	**Top-K**	**High-income**		**Low-income**		**p-value**
		mean	median	mean	median	
English	298	4.44	3.97	4.69	4.17	0.075
Spanish	40	3.62	3.35	2.62	2.67	0.045
	217	3.86	3.35	2.89	2.63	0.000
French	60	4.68	3.84	3.88	3.03	0.059
Russian	40	7.35	6.63	4.67	4.54	0.000
	337	6.12	5.53	4.51	4.54	0.000
Portuguese	295	3.84	3.14	2.61	2.51	0.002
Japanese	387	4.22	3.88	2.46	2.51	0.006
German	87	4.47	3.87	5.38	5.45	0.070
Turkish	141	4.88	4.49	2.60	2.37	0.007

Table 3.13: Significance test to check whether **price** (adjusted to Big Mac price per nationality) in Preply for high and low-income nationalities come from the same distribution. Results with p-value≤0.1 and at least 10 samples per group are represented. Hypothesis test conducted using Kolmogorov-Smirnov test. The disadvantaged group has been highlighted. Languages sorted by decreasing average number of teachers and separator after TOP-LANGUAGES.

Language	**Top-K**	**High-income**		**Low-income**		**p-value**
		mean	median	mean	median	
English	40	3.33	3.00	4.03	4.17	0.011
	486	4.07	3.61	5.40	5.45	0.000
Spanish	40	4.86	4.47	3.38	3.21	0.004
	282	4.35	3.99	3.40	3.21	0.000
French	177	4.61	3.97	5.76	5.67	0.007
Russian	424	6.83	6.08	5.82	5.90	0.000
Portuguese	250	4.04	3.71	2.69	2.01	0.048
Turkish	163	5.68	5.48	4.23	4.37	0.026

Table 3.14: Significance test to check whether **price** (adjusted to Big Mac price per nationality) in Verbling for high and low-income nationalities come from the same distribution. Results with p-value≤0.1 and at least 10 samples per group are represented. Hypothesis test conducted using Kolmogorov-Smirnov test. The disadvantaged group has been highlighted. Languages sorted by decreasing average number of teachers and separator after TOP-LANGUAGES.

Language	**Top-K**	**High-income**		**Low-income**		**p-value**
		mean	median	mean	median	
English	410	4.56	4.24	6.65	6.33	0.000
Spanish	256	5.19	5.21	4.12	4.01	0.000
Russian	82	9.06	8.29	6.61	6.35	0.029

3.2.4 Gender and nationality

Finally, we conduct an analysis on gender and nationality altogether to check if there is a relation between them. This time, we just evaluate the average price using four groups: High-income|Men, High-income|Women, Low-income|Men, and Low-income|Women. Kolmogorov-Smirnov Test cannot be conducted using more than two samples. Nevertheless, a qualitative analysis on mean price is enough to check if the patterns previously analyzed for gender hold for all income levels and vice versa.

Since we create four groups, we take top-100 teachers so as to have a representative sample for all of them. Therefore, we only consider TOP-LANGUAGES for this validation. Results for Italki (see Table 3.15), Preply (see Table 3.16) and Verbling (see Table 3.17) are quite consistent with previous insights. In most cases, women have higher fees for lessons no matter their nationality. Thus, income level does not seem to have a direct effect on how males and females behave on platforms.

Table 3.15: Average price for lessons on Italki with respect to gender and nationality. Only TOP-LANGUAGES were considered.

Language	High-Income		Low-Income	
	Men	**Women**	**Men**	**Women**
English	17.37	17.66	15.90	13.03
Spanish	12.61	14.41	9.38	10.01
French	17.08	19.74	10.30	13.88
Italian	16.30	16.00	-	12.50
Russian	12.12	12.57	6.58	11.25
Portuguese	14.29	18.57	9.47	10.90
Arabic	15.00	50.00	9.97	11.28
Japanese	19.40	15.76	-	-
German	22.73	23.75	8.36	14.50
Korean	14.32	15.77	-	-
Turkish	9.44	10.61	7.18	11.30

Table 3.16: Average price for lessons on Preply with respect to gender and nationality. Only TOP-LANGUAGES were considered.

Language	High-income		Low-income	
	Men	**Women**	**Men**	**Women**
English	15.96	16.30	11.00	11.71
Spanish	15.21	16.35	8.00	11.00
French	20.75	24.93	15.33	19.07
Chinese	16.33	18.24	-	17.00
Italian	18.10	18.51	-	15.00
Russian	17.06	15.47	14.33	16.19
Portuguese	15.79	15.00	11.55	13.00
Arabic	13.33	12.00	11.28	12.64
Japanese	18.50	17.81	-	-
German	27.59	27.50	28.80	15.71
Korean	17.47	18.61	-	-
Turkish	11.73	12.43	8.50	10.57

Table 3.17: Average price for lessons on Verbling with respect to gender and nationality. Only TOP-LANGUAGES were considered.

Language	High-income		Low-income	
	Men	**Women**	**Men**	**Women**
English	22.96	25.46	12.50	15.00
Spanish	16.79	17.45	16.08	15.02
French	24.10	27.10	14.86	15.50
Chinese	23.19	20.99	-	-
Italian	19.00	21.55	-	-
Russian	15.55	16.54	16.33	15.21
Portuguese	22.00	18.71	13.96	15.60
Arabic	14.33	21.00	12.54	14.08
Japanese	20.92	25.17	-	13.00
German	29.31	35.89	-	-
Korean	22.43	24.74	-	-
Turkish	17.90	18.69	-	-

3.2.5 Discussion

The goal of this section is to study whether the price that teachers choose for their lessons might be different depending on protected attributes. This is relevant because although rankings are fair, protected users may still be economically harmed when offering their services. Firstly, we measured whether prices for women and men come from the same distribution across platforms and languages. The results put forward that, whenever there is a significant difference, women always charge more money for their lessons. Secondly, when considering nationality as the protected attribute, the protected group becomes affected. If there is a statistical difference, low-income countries teachers earn less money for their lessons. Nonetheless, since living expenses vary across countries, we established the Big Mac Index as a common framework to consider purchasing power parity. Finally, after converting the price to the common unit (Big Macs), there are fewer languages in which discrimination is spotted. Moreover, in some scenarios,

high-income teachers become the disadvantaged group. This is especially relevant in English, the most popular language, where low-income teachers earn significantly more than those coming from high-income countries on all platforms. In the second most popular language, Spanish, teachers from low-income countries are still harmed across websites.

Finally, we made a qualitative analysis to assess if there was a relation between nationality and gender. Average price was computed for all income level and gender combinations in TOP-LANGUAGES. Results show that, overall, no matter teachers' nationality women have higher fees for their lessons. Thus, the country of origin does not have an impact on how people from different genders choose their price.

3.3 Modeling ranking

As presented at the beginning of this chapter, teachers are sorted according to a black-box function $Score(q, d_i)$ which computes relevance of each teacher d_i for a given query q. In this section, we try to build a function $Score'(q, d'_i)$ that given the language, q, and the publicly available information for teachers, d'_i, is able to approximate $Score(q, d_i)$ accurately. This approximation could let us obtain further information on how ranking is performed on platforms.

As a first step, relevant features were extracted from available datasets to build d'_i. They are different across platforms depending on the data provided by each source. Then, additional features that could be relevant for ordering were inferred. For instance, we included whether a teacher speaks English, and the first letter in their names. Features for all platforms are described in Table 3.18, and Table 3.19 displays which features are available

for each source.

Before modeling ranking using a regression algorithm, statistical correlation between variables was explored. We are especially interested in features correlated with price and position. In Italki (see Table 3.20), *price* is positively correlated with teachers' income level and whether they are featured in the ranking. However, *position* seems to be uncorrelated with other variables. In Preply (see Table 3.21), *price* is again correlated with income level and position is uncorrelated with remaining features. Finally, in Verbling (see Table 3.22), *price* correlation with income level is also found.

Lastly, we tried to model $Score'(q, d'_i)$ using machine learning techniques. At first, we implemented a simple Linear Regression [69] because it can be easily understood and features' relevance can be extracted directly from parameters. Then, we trained a more sophisticated XGBoost Regressor [18] that could improve accuracy. We consider two main approaches for the training process: a unique model used for all languages, and an algorithm trained on each language separately. We train our approximators on English, which is the language with the largest dataset available, and on the whole collection.

At test time, the highest R^2 score obtained among platforms is 0.44. Metrics for Italki (see Table 3.23), Preply (see Table 3.24) and Verbling (see Table 3.25) show that ranking function cannot be reproduced accurately using only publicly available data. Additional private information such as user activity, last connection, etc. might be relevant for the process.

Table 3.18: Attributes considered for position regression across platforms. Which are available for each website can be found in Table 3.19

Field	Description
num_languages	Number of languages spoken by the teacher
speaks_english	Boolean that determines if the teacher speaks english
gender_tuned	Teacher inferred gender
income_level	Teacher origin country GDP per capita
num_teaches	Number of languages taught by a teacher
num_sessions	Number of lessons given by the teacher
is_featured	Teacher is featured in the ranking
avg_lessons_per_students	Average number of lessons per students in the platform
avg_rating	Average rating in the platform
num_ratings	Number of ratings in the platform
price	Price per session
first_letter	Numerical mapping between 0 and 1 for the first letter of the teacher name (A=0, Z=1)
position	Absolute position in raking where 1 is the first teacher in the list. Target for the prediction.

3.3.1 Discussion

In this last analysis, we tried to reproduce ranking algorithms using public data. If a good approximation is achieved, features importance could be derived to get a better understanding of how ordering is performed. However, our metrics show that an accurate simulation cannot be achieved since the highest R^2 score obtained is 0.44 (see Table 3.25). We can conclude that private information, such as usage statistics or last connection, is also relevant for sorting the results. In all platforms, we found a positive correlation

Table 3.19: Attributes available for each plaftorm. Attributes definition can be found in Table 3.18. Missing attributes have been highlighted.

Field	Italki	Preply	Verbling
num_languages	✓	✓	✓
speaks_english	✓	✓	✓
gender_tuned	✓	✓	✓
income_level	✓	✓	✓
num_teaches	✓	✓	✗
num_sessions	✓	✗	✗
is_featured	✓	✓	✗
avg_lessons_per_students	✗	✗	✓
avg_rating	✓	✓	✓
num_ratings	✗	✓	✓
price	✓	✓	✓
first_letter	✓	✓	✓
position	✓	✓	✓

Table 3.20: Correlation matrix for attributes in Table 3.18 considering all languages for Italki.

Variable	1	2	3	4	5	6	7	8	9	10	11
1. num_languages	1.00										
2. speaks_english	0.10	1.00									
3. gender_tuned	-0.08	0.01	1.00								
4. income_level	0.13	-0.01	-0.06	1.00							
5. num_teaches	-0.04	0.04	-0.08	-0.18	1.00						
6. number_sessions	0.05	0.03	-0.01	0.07	0.08	1.00					
7. rating	0.00	0.01	0.00	-0.04	0.06	0.07	1.00				
8. price	0.14	0.02	0.04	0.47	-0.03	0.19	-0.01	1.00			
9. is_pro	0.01	0.01	0.08	0.05	0.06	0.24	0.02	0.39	1.00		
10. position	-0.01	0.01	0.07	0.12	0.03	0.06	0.02	0.05	0.07	1.00	
11. first_letter	-0.03	-0.02	0.04	-0.01	-0.04	-0.02	0.00	0.00	-0.01	-0.01	1.00

Table 3.21: Correlation matrix for attributes in Table 3.18 considering all languages for Preply.

Variable	1	2	3	4	5	6	7	8	9	10	11
1. num_languages	1.00										
2. speaks_english	0.25	1.00									
3. gender_tuned	-0.04	-0.03	1.00								
4. income_level	0.00	-0.01	-0.05	1.00							
5. num_teaches	0.40	0.03	-0.03	-0.10	1.00						
6. num_ratings	0.11	0.06	-0.07	-0.11	0.31	1.00					
7. avg_rating	0.10	0.10	0.00	-0.04	0.15	0.34	1.00				
8. price	0.12	0.05	0.06	0.39	0.06	0.10	0.10	1.00			
9. is_featured	0.03	0.02	0.02	-0.01	0.00	0.09	0.05	0.06	1.00		
10. first_letter	-0.05	-0.07	0.01	0.02	-0.03	-0.02	-0.02	-0.02	0.00	1.00	
11. position	-0.04	-0.07	0.01	0.06	0.01	-0.06	-0.12	-0.05	-0.13	-0.04	1.00

Table 3.22: Correlation matrix for attributes in Table 3.18 considering all languages for Verbling.

Variable	1	2	3	4	5	6	7	8	9	10
1. num_languages	1.00									
2. speaks_english	0.30	1.00								
3. gender_tuned	-0.05	-0.02	1.00							
4. income_level	-0.03	-0.03	-0.06	1.00						
5. avg_lessons_per_students	0.05	0.01	0.02	-0.06	1.00					
6. num_ratings	0.04	0.00	-0.05	0.09	0.19	1.00				
7. avg_rating	0.06	-0.02	0.00	0.06	0.31	0.16	1.00			
8. price	0.17	0.00	0.10	0.36	0.11	0.10	0.05	1.00		
9. first_letter	0.01	-0.03	0.01	-0.01	-0.01	-0.02	-0.03	0.00	1.00	
10. position	-0.10	-0.01	-0.06	0.11	-0.05	-0.02	-0.03	-0.10	-0.03	1.00

Table 3.23: Regression metrics for position on Italki.

Model	R2 Score	Languages considered
Linear Regression	0.02	All
XGBoost Regressor	0.12	All
Linear Regression	0.04	English
XGBoost Regressor	-0.35	English

Table 3.24: Regression metrics for position on Preply.

Model	R2 Score	Languages considered
Linear Regression	0.02	All
XGBoost Regressor	0.29	All
Linear Regression	-0.01	English
XGBoost Regressor	-0.07	English

Table 3.25: Regression metrics for position on Verbling.

Model	R2 Score	Languages considered
Linear Regression	0.02	All
XGBoost Regressor	0.44	All
Linear Regression	0.05	English
XGBoost Regressor	-0.1	English

(around 0.4) between price and income level for teachers.

Chapter 4

Conclusion

Evaluating automated processes that may harm individuals is essential to ensure equal opportunities for everyone in the digital world. For this reason, we decided to evaluate fairness in online tutoring platforms. The goal was to find whether people belonging to protected groups were less likely to appear in the first positions and therefore, to be hired as teachers. Also, we analyzed statistical differences in price for these groups against the overall population. Protected groups considered for the analysis were women and people from low-income countries.

Firstly, results when measuring fairness in rankings (see Section 3.1) put forward that most of the listings are fair across platforms. However, there were some languages for which protected groups were underrepresented in the first positions. *Italki* is the website showing most of these situations. This might mean that fairness is not ensured by platforms when creating rankings. One of the most worrisome imbalances is produced when considering income as the protected attribute. Although there is only one unfair ranking across languages and platforms (66 combinations), it is for the most popular language (English). This is relevant because since it is the language

containing the largest number of teachers and customers, it is the one with the greatest potential for harming individuals.

Secondly, we examined statistical price differences (see Section 3.2). We try to find whether teachers might be choosing different prices based on their self-perception or visitors' behavior. Thus, we evaluate whether prices for protected groups and the remaining population come from the same distribution. When nationality is used for comparison, we found out that teachers from low-income countries have always lower prices than those coming from high-income countries. On the other hand, women have higher fees for all languages on all platforms when there exists a difference. To assess the potential effect of nationality on gender, we computed the average price for all gender and income level combinations. Results showed that, in general, women have higher fees regardless of their nationality. Since living expenses are not equal across countries, we decided to establish a common framework for better comparison. We convert price in dollars to the number of Big Macs someone could buy in their origin country using a lesson fee. This allows comparing price taking into account purchasing power parity. After transformation, there are fewer languages in which we find a significant price difference. Furthermore, low-income teachers become benefited in some languages. Now, it is worth mentioning that English low-income teachers earn significantly more when accounting for purchase parity using the Big Mac index.

Finally, we tried to simulate ranking functions to determine which were the most relevant features for ordering. However, accurate reproduction of the algorithms cannot be achieved using public information according to the obtained metrics. Hence, we can conclude that private data such as usage

statistics are relevant in the process.

Chapter 5

Final remarks

5.1 Limitations and future work

The main limitation of this work is measuring fairness only on the ranking function. Visitors' behavior may be also be a bias source because they only consider a certain group of teachers as eligible due to personal prejudices. For instance, previous research shows that white hosts are more likely to be chosen in AirBnB given *taste-based discrimination* by visitors [27].

This limitation opens a new research topic that requires further data which is not publicly available. Using the number of clicks and bookings each user gets on the platform depending on their ranking position would allow measuring to which extent their number of clients is determined by their position, and how relevant are visitors biases against certain groups.

Besides, we noticed that listings change very often. Given the high computational requirements for crawling all platforms, and the available resources, we worked on a static listing at a given point in time. Nevertheless, replicating this analysis over time could put forward new insights. Another

line of investigation could measure randomness in the rankings and assess whether these variations are fair for all considered groups.

Finally, another variable that could enrich the analysis is where are the teachers located. Our work assumes that teachers live in their origin country, which in many cases, might not be true. With this information, price analysis could be performed for a certain location based on protected attributes.

5.2 Ethical considerations

We conducted a risk assessment for our work in order to identify the parties that could be harmed. This section discusses ethical considerations, risks, and the implemented mitigation strategies. We identify two main groups to pay attention to: users and platforms.

Users

Preserving privacy is the main concern. We believe the risk for teachers is minimal because we use public information anyone can access. However, we infer gender without explicit consent. European legislation with respect to gender is vague and it is not considered sensitive. To preserve privacy and to prevent a person from being offended by the assignment of a certain gender, we release anonymized data that may not lead to identify an individual. As few attributes as possible are released to replicate the results. Moreover, features that might be unique such as price are rounded. Further information about data and transformations is presented in reproducibility Appendix E.

Finally, we think that users' data is analyzed for public interest and

any finding in this work will be utilized to their advantage. At no point information has been used for commercial purposes.

Platforms

Although web scraping is not always permitted by companies, we think that obtaining information is valuable for public interest. We do it in a responsible way to ensure that it never poses a risk to the availability of their website. Since crawling rules were not specified by platforms themselves, we established conservative waiting times to avoid any type of overload in the servers. More details on scraping strategies can be found in Section 2.1.

Bibliography

[1] *Kolmogorov–Smirnov Test*, pages 283–287. Springer New York, New York, NY, 2008.

[2] Andrew Altman. Discrimination. *Stanford Encyclopedia of Philosophy*, forthcoming.

[3] Sihem Amer-Yahia, Shady Elbassuoni, Ahmad Ghizzawi, Ria Mae Borromeo, Emilie Hoareau, and Philippe Mulhem. Fairness in Online Jobs: A Case Study on TaskRabbit and Google. In *International Conference on Extending Database Technologies (EDBT)*, Copenhagen, Denmark, 2020.

[4] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. *Machine Bias*. ProPublica, 2016.

[5] Ricardo A. Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., USA, 1999.

[6] Solon Barocas and Andrew D. Selbst. Big data's disparate impact. *California Law Review*, 104(3):671–732, 2016.

[7] Reuben Binns. On the apparent conflict between individual and group fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 514–524, New York, NY, USA, 2020. Association for Computing Machinery.

[8] Robert J Blake. Current trends in online language learning. *Annual review of applied linguistics*, 31(1):19–35, 2011.

[9] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. NIPS'16, page 4356–4364, Red Hook, NY, USA, 2016. Curran Associates Inc.

[10] N. Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, reprint edition, 2016.

[11] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

[12] David Márquez Carreras and Olga Julià de Ferran. *Un primer curs d'estadística*, volume 48. Edicions Universitat Barcelona, 2011.

[13] Claude Castelluccia and Daniel Le MétayerPanel. *Understanding algorithmic decision-making: Opportunities and challenges*. Panel for the Future of Science and Technology, 2019.

[14] L. Elisa Celis, Anay Mehrotra, and Nisheeth K. Vishnoi. Interventions for ranking in the presence of implicit bias. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 369–380, New York, NY, USA, 2020. Association for Computing Machinery.

[15] Kai-Wei Chang, Vinod Prabhakaran, and Vicente Ordonez. Bias and fairness in natural language processing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): Tutorial Abstracts*, Hong Kong, China, 2019. Association for Computational Linguistics.

[16] Daniel L Chen. This morning's breakfast, last night's game: Detecting extraneous factors in judging. 2016.

[17] Le Chen, Ruijun Ma, Anikó Hannák, and Christo Wilson. Investigating the impact of gender on rank in resume search engines. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–14, New York, NY, USA, 2018. Association for Computing Machinery.

[18] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International*

Conference on Knowledge Discovery and Data Mining, KDD '16, page 785–794. Association for Computing Machinery, 2016.

[19] The European Commission. *2019 Report on equality between women and men in the EU*. The European Commission, 2019.

[20] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. KDD '17, page 797–806, New York, NY, USA, 2017. Association for Computing Machinery.

[21] Bruce Croft, Donald Metzler, and Trevor Strohman. *Search Engines: Information Retrieval in Practice*. Addison-Wesley Publishing Company, USA, 1st edition, 2009.

[22] Shai Danziger, Jonathan Levav, and Liora Avnaim-Pesso. Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences*, 108(17):6889–6892, 2011.

[23] J. Dastin. Amazon scraps secret AI recruiting tool that showed bias against women, 2018.

[24] Thomas H Davenport and Rajeev Ronanki. Artificial intelligence for the real world. *Harvard business review*, 96(1):108–116, 2018.

[25] Miguel Delgado-Rodriguez and Javier Llorca. Bias. *Journal of Epidemiology & Community Health*, 58(8):635–641, 2004.

[26] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, page 214–226, New York, NY, USA, 2012. Association for Computing Machinery.

[27] Benjamin G Edelman and Michael Luca. Digital discrimination: The case of airbnb. com. *Harvard Business School NOM Unit Working Paper*, (14-054), 2014.

[28] Mohamed El Mohadab, Belaid Bouikhalene, and Said Safi. Predicting rank for scientific research papers using supervised learning. *Applied Computing and Informatics*, 15(2):182–190, 2019.

[29] Shady Elbassuoni, Sihem Amer-Yahia, and Ahmad Ghizzawi. Fairness of scoring in online job marketplaces. *ACM/IMS Trans. Data Sci.*, 1(4), 2020.

[30] European Union. *Charter of Fundamental Rights of the European Union*, volume 53. European Union, Brussels, 2010.

[31] Sara Galindo. Evaluating potential biases in commercial people search engines. Master's thesis, Pompeu Fabra University, 2019.

[32] Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. Fairness-aware ranking in search recommendation systems with application to linkedin talent search. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, KDD '19, page 2221–2231, New York, NY, USA, 2019. Association for Computing Machinery.

[33] Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, 38(3):50–57, 2017.

[34] Alexander R Green, Dana R Carney, Daniel J Pallin, Long H Ngo, Kristal L Raymond, Lisa I Iezzoni, and Mahzarin R Banaji. Implicit bias among physicians and its prediction of thrombolysis decisions for black and white patients. *Journal of general internal medicine*, 22(9):1231–1238, 2007.

[35] Lucia-Mihaela Grosu-Rădulescu and Veronica-Maria Stan. Second language acquisition via virtual learning platforms: A case study on romanian experiences. *Rupkatha Journal on Interdisciplinary Studies in Humanities*, 10(3), 2018.

[36] David J Hand and Niall M Adams. Data mining. *Wiley StatsRef: Statistics Reference Online*, pages 1–7, 2014.

[37] Fariba Karimi, Claudia Wagner, Florian Lemmerich, Mohsen Jadidi, and Markus Strohmaier. Inferring gender from names on the web: A comparative evaluation of gender detection methods. In *Proceedings of the 25th International Conference Companion on World Wide Web*, WWW '16 Companion, page 53–54, Republic and Canton of Geneva, CHE, 2016. International World Wide Web Conferences Steering Committee.

[38] Fariba Karimi, Claudia Wagner, Florian Lemmerich, Mohsen Jadidi, and Markus Strohmaier. Inferring gender from names on the web: A comparative evaluation of gender detection methods. WWW '16 Companion, page 53–54, Republic and Canton of Geneva, CHE, 2016. International World Wide Web Conferences Steering Committee.

[39] Marzieh Karimi-Haghighi and Carlos Castillo. *Efficiency and Fairness in Recurring Data-Driven Risk Assessments of Violent Recidivism*, page 994–1002. Association for Computing Machinery, New York, NY, USA, 2021.

[40] Brendan F. Klare, Mark J. Burge, Joshua C. Klontz, Richard W. Vorder Bruegge, and Anil K. Jain. Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security*, 7(6):1789–1801, 2012.

[41] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human Decisions and Machine Predictions*. *The Quarterly Journal of Economics*, 133(1):237–293, 08 2017.

[42] Alexandros Labrinidis and H. V. Jagadish. Challenges and opportunities with big data. *Proc. VLDB Endow.*, 5(12):2032–2033, 2012.

[43] Tie-Yan Liu. Learning to rank for information retrieval. *Found. Trends Inf. Retr.*, 3(3):225–331, 2009.

[44] Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. *Gender Bias in Neural Natural Language Processing*, pages 189–202. Springer International Publishing, Cham, 2020.

[45] Emmie Matsuno and Stephanie L Budge. Non-binary/genderqueer identities: A critical review of the literature. *Current Sexual Health Reports*, 9(3):116–120, 2017.

[46] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *CoRR*, abs/1908.09635, 2019.

[47] Ted R Mikuls, Kenneth G Saag, Varghese George, Amy S Mudano, and Samprit Banerjee. Racial disparities in the receipt of osteoporosis related healthcare among community-dwelling older women with arthritis and previous fracture. *The Journal of rheumatology*, 32(5):870–875, 2005.

[48] Todd D Nelson. *Handbook of prejudice, stereotyping, and discrimination*. Psychology Press, 2009.

[49] Mei Ngan and Patrick Grother. Face recognition vendor test (frvt) - performance of automated gender classification algorithms, 2015.

[50] Ziad Obermeyer and Ezekiel J Emanuel. Predicting the future—big data, machine learning, and clinical medicine. *The New England journal of medicine*, 375(13):1216, 2016.

[51] Council of Europe. *European Convention on Human Rights*. 1950.

[52] C. O'Neil. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown, reprint edition, 2017.

[53] Li Lian Ong. *The Economics of the Big Mac Standard*, pages 51–87. Palgrave Macmillan UK, London, 2003.

[54] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, 1999. Previous number = SIDL-WP-1999-0120.

[55] Jiaul H. Paik. A novel tf-idf weighting scheme for effective ranking. SIGIR '13, page 343–352, New York, NY, USA, 2013. Association for Computing Machinery.

[56] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, page 560–568, New York, NY, USA, 2008. Association for Computing Machinery.

[57] Sharese N. Porter. *Poverty, Discrimination, and Health*, pages 23–53. Springer International Publishing, Cham, 2019.

[58] Nunung Nurul Qomariyah, Dimitar Kazakov, and Ahmad Nurul Fajar. Predicting user preferences with xgboost learning to rank method. In *2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, pages 123–128, 2020.

[59] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. Mitigating bias in algorithmic hiring: Evaluating claims and practices.

In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 469–481, New York, NY, USA, 2020. Association for Computing Machinery.

[60] S. Russell. *Human Compatible: Artificial Intelligence and the Problem of Control*. Penguin Books, reprint edition, 2020.

[61] S.J. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 2 edition, 2002.

[62] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.*, 24(5):513–523, 1988.

[63] Sefik Ilkin Serengil and Alper Ozpinar. Lightface: A hybrid deep face recognition framework. In *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 23–27. IEEE, 2020.

[64] Leiyu Shi, Jenna Tsai, Patricia Collins Higgins, and Lydie A Lebrun. Racial/ethnic and socioeconomic disparities in access to care and quality of care for us health center patients compared with non–health center patients. *The Journal of ambulatory care management*, 32(4):342–350, 2009.

[65] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815*, 2017.

[66] Songül Tolan. Fair and unbiased algorithmic decision making: current state and future challenges. *arXiv preprint arXiv:1901.04730*, 2019.

[67] Michelle Van Ryn and Jane Burke. The effect of patient race and socio-economic status on physicians' perceptions of patients. *Social science & medicine*, 50(6):813–828, 2000.

[68] Ho Chung Wu, Robert Wing Pong Luk, Kam Fai Wong, and Kui Lam Kwok. Interpreting tf-idf term weights as making relevance decisions. *ACM Trans. Inf. Syst.*, 26(3), 2008.

[69] Xin Yan and Xiao Gang Su. *Linear Regression Analysis: Theory and Computing*. World Scientific Publishing Co., Inc., USA, 2009.

[70] Ke Yang and Julia Stoyanovich. Measuring fairness in ranked outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*, SSDBM '17, New York, NY, USA, 2017. Association for Computing Machinery.

[71] Ke Yang, Julia Stoyanovich, Abolfazl Asudeh, Bill Howe, HV Jagadish, and Gerome Miklau. A nutritional label for rankings. In *Proceedings of the 2018 International Conference on Management of Data*, SIGMOD '18, page 1773–1776, New York, NY, USA, 2018. Association for Computing Machinery.

[72] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. Fa*ir: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, page 1569–1578, New York, NY, USA, 2017. Association for Computing Machinery.

[73] Meike Zehlike, Ke Yang, and Julia Stoyanovich. Fairness in ranking: A survey. *CoRR*, abs/2103.14000, 2021.

[74] Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. Deep learning based recommender system: A survey and new perspectives. *ACM Comput. Surv.*, 52(1), 2019.

Appendix A

Languages available per platform

Table A.1: Number of teachers for languages selected for analysis in the considered platforms. Sorted by descending average number of teachers across websites. Languages highlighted in green will be considered as TOP-LANGUAGES because they have at least 100 teachers in average.

Language	Italki	Preply	Verbling	Average
English	4399	11693	729	5607.00
Spanish	1841	4650	554	2348.33
French	823	1506	168	832.33
Chinese	967	1208	135	770.00
Italian	517	1073	147	579.00
Russian	617	966	98	560.33
Portuguese	360	939	89	462.67
Arabic	506	647	103	418.67
Japanese	593	580	58	410.33
German	498	593	75	388.67
Korean	239	240	46	175.00
Turkish	163	253	38	151.33
Polish	132	118	11	87.00
Hindi	96	124	14	78.00
Dutch	83	119	19	73.67
Greek	83	95	24	67.33
Indonesian	65	93	18	58.67
Persian	117	23	23	54.33
Thai	79	57	19	51.67
Vietnamese	95	35	19	49.67
Hebrew	65	65	10	46.67
Romanian	42	16	13	23.67

Table A.2: Number of teachers for languages that were not selected for analysis. Sorted by descending average number of teachers across websites.

Language	Italki	Preply	Verbling	Average
Ukrainian	121	91	2	71.33
Serbian	106	77	6	63.00
Tagalog	80	97	2	59.67
Urdu	37	76	6	39.67
Hungarian	57	36	7	33.33
Swedish	34	52	6	30.67
Czech	52	21	2	25.00
Catalan	65	2	3	23.33
Croatian	48	14	5	22.33
Punjabi	24	26	4	18.00
Tamil	17	24	3	14.67
Bengali	18	17	3	12.67
Latin	15	22	0	12.33
Danish	20	12	4	12.00
Bulgarian	19	12	4	11.67
Slovak	26	5	2	11.00
Norwegian	14	13	4	10.33
Finnish	11	7	1	6.33
Gaelic	8	0	0	2.67
Icelandic	5	1	1	2.33

Appendix B

Italki languages grouping

In order to match languages in Verbling and Preply, I needed to group several languages in Italki into a higher level. In this table, I present the matching between languages considered for analysis and languages as they are collected in Italki.

Table B.1: Grouped languages in Italki to match other platforms.

Language	Grouped languages
Chinese	Chinese (Mandarin), Chinese (Cantonese), Chinese (Hakka), Chinese (Hokkien), Chinese (Other), Chinese (Shanghainese), Chinese (Taiwanese)
Arabic	Arabic, Arabic (Egyptian), Arabic (Gulf), Arabic (Levantine), Arabic (Maghrebi), Arabic (Modern Standard), Arabic(Sudanese)
Japanese	Japanese, Japanese (Okinawan), Japanese Sign Language
Greek	Greek, Greek (Ancient)
Gaelic	Gaelic (Irish), Gaelic (Manx), Gaelic (Scottish)

Appendix C

Nationalities in Preply

Table C.1: Number of unique teachers per nationality in Preply. Nationalities are represented using their corresponding ISO codes.

Nationality	Teachers	Nationality	Teachers	Nationality	Teachers
FR	509	AU	20	AE	3
IT	499	DZ	19	BH	3
CN	434	HK	16	MK	3
RU	376	KZ	16	PA	3
JP	310	MY	16	PR	3
BR	257	PE	16	SN	3
DE	241	CD	14	ZW	3
UA	230	PH	14	AF	2
ES	203	CL	13	GH	2
EG	189	SA	13	LV	2
TR	176	BA	12	LY	2
KR	171	HT	12	MO	2
US	166	AZ	11	MU	2
GB	150	DO	11	MZ	2
IN	133	MD	11	NI	2
PL	133	AL	10	PY	2
CA	109	AM	10	SI	2
PT	102	EC	10	SK	2
CO	93	HR	10	ZM	2
GR	89	IQ	10	BI	1
ID	88	NZ	9	BM	1
MX	87	TM	9	BZ	1
VE	77	YE	9	CG	1
ZA	74	CR	8	CV	1
NL	71	KG	8	ET	1
AR	64	NG	8	FI	1
TW	63	CI	7	GA	1
CM	57	CY	7	GN	1
TH	57	CZ	7	GQ	1
PS	51	GE	7	HN	1
MA	49	IE	7	KW	1
LB	46	BD	6	LT	1
BE	44	ME	6	MG	1
IL	44	UY	6	NC	1
CH	39	UZ	6	NP	1
RS	36	GT	5	OM	1
AT	33	HU	5	PF	1
VN	31	SE	5	QA	1
BY	30	AO	4	RE	1
PK	26	BG	4	RW	1
TN	23	BJ	4	SV	1
JO	22	KE	4	SY	1
RO	21	SG	4	TG	1

Appendix D

Kolmogorov-Smirnov Test

Kolmogorov-Smirnov Test [1] is a non-parametric test used to determine whether two independent empirical samples come from the same distribution. In this work, we consider price for protected teachers and non-protected teachers as two independent samples in our analysis. $X = (x_1, x_2, ..., x_{n_1})$ represents price for teachers belonging to D_{NP}^{lang}, and $Y = (y_1, y_2, ..., y_{n_2})$ for protected teachers within D_P^{lang}.

The test proceeds as follows [12]. First, we compute the empirical distributions F_n for each sample, X and Y, containing n_1 and n_2 i.i.d. observations respectively.

$$F_n^1(z) = \sum_{i=1}^{n_1} \mathbb{1}_{[-\infty,z]}(x_i)$$
$$F_n^2(z) = \sum_{i=1}^{n_2} \mathbb{1}_{[-\infty,z]}(y_i)$$

Then, a D (or Kolmogorov–Smirnov) statistic is computed comparing

both empirical distributions.

$$D = \sup_{z \in \mathbb{R}} |F_n^1(z) - F_n^2(z)|$$

This statistic will be close to 0 if distributions are similar. However, since p-value is the most common metric for statistical tests in literature, we performed a two-sided hypothesis test to standardize our results. Null hypothesis, H_0, is that both distributions are identical. Small p-values represent that such distributions are unlikely to be obtained under the null hypothesis. On the other hand, high p-values indicate high probability of observing that behavior assuming that both distributions are equal.

Appendix E

Reproducibility: code and data

This appendix contains details about how our insights can be reproduced, and the published code and data. All the scripts used for crawling and evaluation can be found in a GitHub repository[1]. It also contains the data required to reproduce the experiments.

However, some slight modifications were made to data to ensure to ensure data anonymization. All the details regarding ethical data handling can be found in Section 5.2.

Notice that if crawling is reproduced, you will require a valid API key from Gender API[2].

Repository structure

The repository is structured in three different folders:

- `/data_acquisition/`. Contains the Python scripts used to crawl the

[1]https://github.com/javirandor/online-tutoring-analysis
[2]https://gender-api.com

information from websites. Notice that sources might be updated and the provided code might be no longer useful.

- `/analysis/`. Jupyter Notebooks required to reproduce results.
- `/data/`. Anonymized datasets.

Released data

As explained in Section 5.2, released data should not disclose any information that may identify an individual to mitigate risks. For this reason, data available in the repository is slightly different to the one used for the analysis, and reproduction might lead to different results.

We released two different ranking files. Those datasets within the folder `simplified_ranking/`, contain listings with protected attributes and language that can be used for FA*IR and statistical analysis. No personal information is provided and price was rounded to the closest multiple of 3 to avoid identification. On the other hand, files within `modeling_ranking/` provide the data to train the regressors for position. Since they use more information that can lead to identification, further anonymization was conducted. We provide the transformed dataset that can be used for training because transformation requires sensitive features such as name. Moreover, language is not provided, and price and first letter are scaled. All notebooks used for the analysis have been slightly modified to take this data as input.