BIAS IN SEARCH: EVALUATING SEARCH RESULTS THROUGH RANK AND RELEVANCE BASED MEASURES

by
GIZEM GEZICI

Submitted to the Graduate Engineering and Natural Sciences
in partial fulfilment of
the requirements for the degree of Doctor of Philosophy

Sabancı University
June 2022

ACKNOWLEDGEMENTS

First and foremost, I am deeply indebted to my advisor Prof. Yücel Saygın who made this endeavor possible. His extensive knowledge, guidance, and motivation encouraged and carried me through all the stages of preparing my thesis. I am also extremely grateful to my thesis co-advisor, Prof. Emine Yılmaz who guided me throughout this journey with her valuable feedback. Additionally, I am also grateful to Asst. Prof. Aldo Lipani whose insight and knowledge especially into the bias analysis in search results steered me through this research. Furthermore, I could not have undertaken this journey without my defense committee, who generously provided knowledge and expertise.

I would like to thank Dr. Stefan Räbiger for the valuable discussions and his patience with my endless questions as well as the greatest celebration cake ever! I would also like to thank Ahmmad O.M. Saleh for the additional experiments he fulfilled to support me during the last weeks of my thesis preparation, he is the best colleague ever! I am also grateful to my best friend Efe Naci Erten for believing in me at every stage of my thesis, during good and bad times and encouraged me to accomplish further with his invaluable spiritual guidance.

Last but not least, I am extremely grateful to my beloved Turgay Gündoğdu for always being by my side, cherishing me with his supporting words and supporting me even by annotating documents to help me for improving the models during the dark pandemic times. Without him, this thesis would not have been possible. Finally, I would like to express my love and gratitude to my parents Şule and Metin Gezici and my brother Barış Gezici, my beloved grandpa Turgut Avcıoğlu and grandma Melek Avcıoğlu. Without their tremendous encouragement and endless questions (!) related to the thesis process in the past few years, it would have been impossible for me to find the motivation to complete this thesis.

ABSTRACT

BIAS IN SEARCH: EVALUATING SEARCH RESULTS THROUGH RANK AND RELEVANCE BASED MEASURES

GIZEM GEZICI

Computer Science and Engineering
Ph.D Dissertation, June 2022

Dissertation Supervisor: Prof. Yücel Saygın
Dissertation Co-Supervisor: Prof. Emine Yılmaz

Keywords: search, bias, stance bias, ideological bias, gender bias, controversial issues, online education

Search is ubiquitous. People continue to seek information through popular search engines, Bing and Google as well as online search platforms, YouTube. Nonetheless, they tend to think that these platforms are objective by only displaying information without injecting any bias. Since users are more susceptible to bias when they are unaware of it, it is important to evaluate the retrieved search results of the aforementioned platforms with respect to bias. This thesis analyses two main things as search engine bias towards controversial issues and gender bias in the context of online education. For evaluating specifically search engine bias, three novel rank and relevance-based measures have been proposed and search results of two widely-used search engines Google and Bing have been analysed through web documents' content with respect to stance (in support or against), and ideological bias (conservative or liberal). Then, the impact of geolocation on the bias has been investigated. Lastly, in the scope of search engine bias, the source of bias has been tracked, to check whether the bias (if exists) comes from the input data, or the ranking algorithm. For assessing gender bias in online education, two new rank and relevance based measures that are more suitable in the scope of gender bias have been proposed. Further, video search results returned by YouTube towards the queries in STEM and NON-STEM fields have been analysed using narrators' information. Lastly, the source of gender bias has been investigated by proposing the specifically-curated

gender bias measures.

ÖZET

ARAMA PLATFORMLARINDA ÖN YARGI: ARAMA SONUÇLARININ SIRALAMA VE İLGİLİLİK TEMELLI METRİKLER İLE DEĞERLENDİRİLMESİ

GIZEM GEZICI

Bilgisayar Bilimi ve Mühendisliği Doktora Tezi, Haziran 2022

Tez Danışmanı: Prof. Dr. Yücel Saygın
Tez Eş-Danışmanı: Prof. Dr. Emine Yılmaz

Anahtar Kelimeler: arama, ön yargı, tutumsal ön yargı, ideolojik ön yargı, cinsiyetçi ön yargı, tartışmalı konular, çevrimiçi eğitim

Arama platformları hayatımızın her yerinde. İnsanlar popüler arama motorları olan Bing ve Google üzerinden olduğu gibi YouTube gibi diğer arama platformlarından da bilgi arayışı içerisindeler. Bununla birlikte, kullanıcılar arama platformlarını hiçbir ön yargı katmadan yalnızca bilgiyi sunan objektif platformlar olarak görüyorlar. Kullanıcılar farkında olmadıklarında ön yargılara karşı daha da savunmasız kalmaları sebebiyle, arama platformlarından dönen sonuçların ön yargı açısından analiz edilmesi önemli. Bu tez esasen arama motorlarında bulunan ön yargıları tartışmalı konular üzerinden, ve arama platformlarında bulunan cinsiyetçi ön yargıları da çevrimiçi eğitim odağında analiz ediyor. Arama motorlarına özgü ön yargıları analiz etmek için, üç yeni sıralama ve ilgililik tabanlı metrik önerildi. Bu metrikler kullanılarak Bing ve Google'ın arama sonuçları web dokümanlarının içeriği üzerinden tutumsal (destekliyor veya karşı) veya ön yargı ve ideolojik (muhafazakar veya liberal) ön yargı olarak iki kısımda incelendi. Ek olarak, lokasyonun bu ön yargı sonuçlarına olan etkisi incelendi. Son olarak, arama motoru sonuçları ön yargının kaynağı – ön yargının veri setinden mi yoksa sıralama algoritmasından mı geldiğinin anlaşılması için incelendi. Cinsiyetçi ön yargının çevrimiçi eğitimde değerlendirilmesi için, bu kapsama uygun iki yeni sıralama ve ilgililik tabanlı metrik önerildi. Sonrasında YouTube'un STEM ve NON-STEM alanları ile ilgili sorgulara karşılık döndürdüğü video arama sonuçları video'daki anlatıcının bilgileri kullanılarak incelendi. Son olarak, video arama sonuçlarındaki cinsiyetçi ön yargının kaynağı bu amaca özgü bir şekilde adapte edilmiş metriklerle araştırıldı.

This thesis is dedicated to my family
For their endless love and support...

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

Search is ubiquitous. According to SmartSights (2018), 46.8 percent of the world's population used the internet in 2017, and that figure is predicted to rise to 53.7 percent by 2021. According to InternetLiveStats (2018), around 3.5 billion Google searches are conducted daily on average. These figures demonstrate that search engines have surpassed traditional broadcast media as a *primary* source of information and have become the 'gatekeepers to the Web' for many individuals, particularly in the last decade. As information seekers conduct more searches on the Web, they are more impacted by Search Engine Result Pages (SERPs) covering a broad range of topics (e.g., work, entertainment, religion, and politics). For example, it is well known that during elections, individuals make repeated Web searches for political candidates and events such as "democratic debate", "Donald Trump", "climate change" (Kulshrestha, Eslami, Messias, Zafar, Ghosh, Gummadi & Karahalios, 2018). The SERPs returned in response to these queries may have an effect on voting decisions, as asserted by Epstein & Robertson (2015), who claim that *manipulated* search rankings might alter indecisive voters' voting preferences by at least 20%.

Apart from the search engine bias through favouring different viewpoints, the search results of other widely-used search platforms need to be audited as well. According to recent studies, YouTube has been declared as the world's second-most visited website and second-most used social platform worldwide (Aslam, 2021). In the third quarter of 2020 during the pandemic, YouTube had roughly 80% market penetration in the UK, outperforming Facebook, WhatsApp, Instagram, and Twitter by number of active users and it had the highest reach among users aged 15 to 25 with 82% of this demographic group (Ceci, 2021). Although YouTube is popular as an entertainment medium, it has become a valuable alternative learning resource to written textual content such as blogs (Chintalapati, Srinivas & Daruri, 2016). Chtouki, Harroud, Khalidi & Bennani (2012) report a study showing that visual explanations help students to understand and remember the complex concepts much better. These studies justify that YouTube has been a widely-used platform as well as an effective tool for improving student's learning and engagement. Although there have been some attempts to mitigate gender bias in instructional materials of traditional education in schools (Center, 2020), the problem might still exist in online educational materials such as blogs, specifically curated online educational platforms, and YouTube videos. It has been observed that among university students, female students tend to be affected more by contexts than male students while searching online information (Zhou, 2014). Thus, analysing gender bias in online educational materials is necessary as well.

What we see for a search query is determined by search platforms. Since many people are exposed to information via online search platforms, it is reasonable to assume that search platforms are objective in terms of different perspectives, or gender representation. Yet, search results might not be impartial, i.e. they do not necessarily provide all viewpoints on a given query, and they can be biased towards a specific perspective or they might favour a specific gender group over another. Since search results are retrieved based on relevancy, which is calculated using numerous features and complex algorithms, for those search platforms providing unbiased search results is not often the primary concern. Therefore, it is essential to examine the search results with respect to bias.

In this thesis, web search bias is evaluated on two popular search engines, Bing and Google with novel bias measures that take into account rank and relevance information and an evaluation framework. Web search bias is interpreted in stance (support or against) and ideological bias (conservative or liberal) towards controversial topics. In addition to search engine bias, gender bias in YouTube is also analysed in the context of educational queries of STEM and NON-STEM majors. For evaluating online gender bias in the educational context, new bias measures that leverage both rank and relevance information are presented to assess whether search results are biased towards a specific gender group, i.e. male or female.

1.1 Web Search Bias Evaluation

Search engines have evolved into an integral component of our daily life. While search engines are commonly used for seeking information, the majority of internet users believe they give *impartial* results, i.e. act as facilitators for accessing information on the Web (Goldman, 2008). There are, however, counter-examples to this belief. A recent dispute between US President Donald Trump and Google exemplifies this, with Mr. Trump accusing Google of displaying only negative news about him when his name is searched, to which Google responded with the following statements: “When users type queries into the Google Search bar, our goal is to ensure they receive the most relevant answers in a matter of seconds” and “Search is not used to advance a political agenda, and we do not bias our results toward any political ideology” (Ginger & David, 2018). We intend to throw some light on that argument in Chapter 3 by not focusing exclusively on Donald Trump-related inquiries but by performing an in-depth examination of search responses to a broad range of contentious subjects using concrete evaluation metrics.

Bias is described in terms of the representational balance of Web documents returned from a database in response to a specific query (Mowshowitz & Kawaguchi, 2002a). When a user submits a search query to a search engine, documents from many sources are collected, ranked, and presented to the user. Assume that a user searches for *2016 presidential election* and is presented with the top-n ranked results. In such a search scenario, the retrieved results may favor certain political perspectives over others, thereby failing to deliver impartial knowledge for the given query, as Mr. Trump asserts, albeit without scientific evidence. Hence, the potential *undue emphasis* of specific perspectives (or viewpoints) in the retrieved results lead to bias (Kulshrestha et al., 2018). With regard to the definition of bias and the scenario described, if there is an imbalanced representation, i.e. skewed or slanted distribution, of opinions in a SERP, i.e. not just in political searches, toward the query's topic, we consider this SERP to be *biased* for the given search query.

In prior works, two different notions of fairness as *individual fairness* and *group fairness* in ranked outputs have been investigated. Individual fairness requires that similar individuals should be treated similarly, whereas group fairness requires that the disadvantaged group be treated similar to the advantaged group or the entire population (Dwork, Hardt, Pitassi, Reingold & Zemel, 2012). Formal definitions of group fairness is composed of *statistical parity*, *demographic parity* or more generally known as *equality of outcome* (Dwork et al., 2012), and *equality of opportunity* (Hardt Google, Price & Srebro, 2016). Note that, throughout this thesis, bias is examined in the context of a specific type of bias known as *statistical parity*, *demographic parity* or, more broadly known as *equality of outcome*, which states that given a population divided into groups, the groups represented in the system's output should be *equally represented*. This is in contrast to the other widely used metric, generally referred to as *equality of opportunity*, which states that given a population divided into groups, the groups in the output should be represented according to their population proportion, or base rates. Since, the *equal* representation is required in an unbiased scenario, *equality of outcome* is a more suitable group fairness measure in the context of this thesis.

Bias is especially critical when the query topic is *controversial* and has conflicting viewpoints; in this case, it becomes even more important that search engines deliver results that are representative of multiple perspectives, which means that they do not favor one viewpoint over another. Otherwise, as with the case of elections, this may have a dramatic effect on the public, resulting in polarisation in society about *controversial* topics. Thus, Chapter 3 focuses on search engine bias in the scope of controversial topics.

In Chapter 3, two new bias measures using rank and relevance information and a new bias evaluation framework are presented. The evaluation framework concentrates on the top-10 SERPs coming from the *news* sources to investigate two major search engines in terms of bias. For this analysis, annotating top-10 news SERPs has been fulfilled via crowd sourcing. The bias analysis is performed on *news* SERPs intentionally, as they often reflect a particular viewpoint on a topic (Alam & Downey, 2014). In the context of this chapter, it is also shown that how the introduced bias evaluation framework can be used to quantify bias in the SERPs of Bing and Google in response to queries of *controversial* topics. The bias analysis is primarily two-fold where stance bias in SERPs is evaluated first, then that evaluation is used as a proxy for quantifying ideological bias expressed in the SERPs of the search engines. Recent studies (99Firms, 2019; Sarcona, 2019) indicate that more than on average 70% of all clicks occur on the first page results; hence the initial web search bias analysis in Chapter 3 focuses solely on the top-10 results to demonstrate the existence of bias, if any.

1.2 The Impact of Geolocation on Search Engine Bias

As a follow-up study, in Chapter 4 the impact of location on search engine bias is investigated using the news search results returned by Bing and Google in the UK and US locations. For this, a similar evaluation procedure that has been proposed in Chapter 3 is applied on the news SERPs of two popular search engines. Further, it is analysed if different locations have an impact on the existence of bias as well as the magnitude of bias in Google and Bing SERPs in the UK and US locations.

1.3 Investigating the Source of Search Engine Bias

In addition to the search engine bias in the top-10 SERPs, and the impact of location on the bias results of search engines, further the source of bias is investigated, i.e. if it comes from the input data, or the ranking algorithm using the whole corpus of retrieved documents for the controversial queries in Chapter 3, as well. Thus, Chapter 5 attempts to track the source of bias using state-of-the-art approaches to establish an automated model to obtain annotations for the retrieved documents in the news channel of two popular search engines. Since crowd-labelling is a costly process to get annotations for the whole corpus, i.e. on average 250 web pages, source of bias analysis requires an automated model as the initial step.

Tracking the source of bias is valuable in the context of bias analysis in search since just returning an imbalanced representation of diverse viewpoints does not constitute evidence of a search engine's ranking algorithm being biased. A skewed SERP could be caused by the corpus itself, i.e. if the documents indexed and returned for a specific topic have a slanted distribution, implying that the ranking algorithm provides a biased result set as a result of the biased corpus. To distinguish algorithmic from corpus bias, one must look at the source of bias in addition to performing a skewed list analysis of the top-n search results.

Nonetheless bias, whether corpus or algorithmic, would contradict with the expectation that an IR system should be fair, accountable, and transparent (Culpepper, Diaz & Smucker, 2018). Additionally, it was observed that individuals are more susceptible to bias when they are unaware of it (Bargh, Gollwitzer, Lee-Chai, Barndollar & Trötschel, 2001). Moreover, Epstein, Robertson, Lazer & Wilson (2017) demonstrated that informing users about bias can be beneficial at mitigating the effect of search engine manipulation effect (SEME). Thus, search engines should inform their users about bias and mitigate the possibility of SEME by increasing their accountability, thereby mitigating the negative impacts of bias and serving only as facilitators, as they usually claim to do. Based on these reasons, throughout this thesis there will be some attempts to measure search engine bias as well as investigate the source of this bias.

1.4 Online Gender Bias Evaluation

Apart from the search engine bias analysis, the second part of the thesis focuses on gender bias, which comes from the societal stereotypes about different gender groups, in search in the context of education. In recent years, since the use of online educational materials for learning have increased, in addition to traditional, analysing gender bias in online learning have become necessary as well.

Stereotypes are defined as beliefs regarding the characteristics, attributes, and behaviors of members of certain groups (Hilton & Von Hippel, 1996). Such beliefs, referring to society's stereotypes, often contain oversimplifications and prejudices about a specific group (Piatek-Jimenez, Cribbs & Gill, 2018). Gender stereotypes begin to develop in early ages and these stereotypes about science, technology, engineering and mathematics (STEM) have severe consequences for motivation towards STEM fields (McGuire, Mulvey, Goff, Irvin, Winterbottom, Fields, Hartstone-Rose & Rutland, 2020). The early emerging gender stereotypes related to STEM are further strengthened in adolescence by the presence of male teachers and gender-imbalanced classrooms in STEM majors (Riegle-Crumb, Moore & Buontempo, 2017). A common stereotype is that STEM careers are for certain social groups such as European or American white males, (Barman, 1997; Bodzin & Gehringer, 2001) and this stereotype might signal to women and racial minority students that their group does not belong and is not successful in the STEM field (Good, Rattan & Dweck, 2012), thereby making them feel less welcoming, more insecure, and less motivated in STEM (London, Rosenthal, Levy & Lobel, 2011). Further, these stereotypes continue in the workplace and broader society, leading to the underrepresentation of woman in STEM fields (Piatek-Jimenez et al., 2018). For instance, in the UK only 22%, and in the US only 24% of the STEM workforce is constituted by women (Noonan, 2017; WISE, 2018).

Gender stereotypes are a common source of bias that emerge when an individual or a group is systematically treated favourably or unfavourably, referring to *individual* or *group fairness* respectively and there is a need to investigate gender representation in educational resources. In fact, the European Institute for Gender Equality states that gender stereotypes still exist in educational materials (EU, 2017). There are some guidelines on how to evaluate diversity in educational materials, for example Michigan in the United States issued a report in 2020 as a guidance for the experts in evaluating instructional materials in terms of bias (Center, 2020). These guidelines contain templates for scorecards to help the experts in their evaluation. Schools may try to implement the suggested guidelines and update/change their educational materials to mitigate the bias, but the problem may still persist since students are increasingly referring to online materials such as blogs, online educational websites, and YouTube videos. Nonetheless, gender seems to be an influential factor even in online search; it has been observed that university female students might be more readily to be affected by contexts than male students during online information seeking (Zhou, 2014). Thus, evaluating bias in online educational materials is very critical as well. Based on these, Chapter 6 focuses on gender bias in online educational videos retrieved by YouTube towards the majors of STEM and NON-STEM fields.

1.5 Investigating the Source of Online Gender Bias

As a follow-up study, Chapter 7 tracks the source of online gender bias in education using the adapted version of the bias measures with the similar evaluation procedure proposed in Chapter 6. Similar to Chapter 5, various state-of-the-art approaches have been leveraged to obtain annotations for the whole corpus of the returned YouTube educational videos, i.e. 200 video search results per query, towards the queries of STEM and NON-STEM fields. Hence, Chapter 7 aims to shed light on online gender bias in the educational materials of YouTube.

1.6 Outline of the Thesis

The rest of the thesis is organised as follows. Chapter 2 gives related work mainly in the research areas of search engine bias and online gender bias in education. The first part of the thesis deals with search engine bias, while the second part concentrates on gender bias in online educational materials. The first part starts with Chapter 3 which evaluates bias in top-10 news SERPs of two major search engines, Bing and Google by proposing two novel bias measures that leverage rank and relevance information in search results as well as an evaluation framework to quantify bias. As a follow-up study, Chapter 4 investigates the effect of different locations on the search engine bias results using the news SERPs of Google and Bing in the UK and US. As the last step of the search engine bias analysis, Chapter 5 tracks the source of bias to check whether the bias (if exists) comes from the data, or the ranking algorithms of the two major search engines. Then, the second part of the thesis starts with Chapter 6 which focuses on gender bias in online learning using the educational videos retrieved by YouTube towards the majors from STEM and NON-STEM fields. As a follow-up study, Chapter 7 aims to investigates the source of online gender bias in the educational YouTube videos. Lastly, Chapter 8 concludes the thesis and provides potential future work.

2. BACKGROUND & RELATED WORK

Bias analysis in search platforms has drawn considerable attention in recent years (Baeza-Yates, 2016; Chen & Yang, 2006; Hannák, Wagner, Garcia, Mislove, Strohmaier & Wilson, 2017; Kay, Matuszek & Munson, 2015; Mowshowitz & Kawaguchi, 2002b; Noble, 2018; Pan, Hembrooke, Joachims, Lorigo, Gay & Granka, 2007; Singh, Chayko, Inamdar & Floegel, 2020; Tavani, 2012) due to concerns that search platforms may manipulate search results or propagate societal stereotypes, thereby influencing users. The primary reason for these worries is that search platforms have evolved into the primary source of information (Dutton, Blank & Groselj, 2013), with Pew (2014) and Reuters (2018) studies showing that more individuals obtain their news from search engines than from social media. Users placed a higher trust on search engines' accuracy (Elisa Shearer, 2018; Newman et al., 2018; Newman, Fletcher, Kalogeropoulos & Nielsen, 2019), and many internet-using adults in the United States even utilize search engines to fact-check information (Dutton, Reisdorf, Dubois & Blank, 2017). In additon to search engines, according to Aslam (2021), YouTube has been reported as the world's second-most visited website and second-most used social platform worldwide in the recent studies. In the third quarter of 2020 during the pandemic, YouTube had roughly 80% market penetration in the UK and the highest reach among users aged 15 to 25 with 82% of this demographic group (Ceci, 2021).

To better understand how this growing use of search platforms and trust in them may have unintended consequences for the public, and what methods might be used to quantify those consequences, the following sections review the research areas related to search engine bias through document stances and ideologies as well as gender bias in education. Specifically, first automatic stance detection which is related to search engine bias, then fair ranking evaluation and finally search bias quantification are reviewed in the scope of search engine bias and gender bias in online education.

2.1 Opinion Mining and Sentiment Analysis

Contrastive Opinion Modeling is a type of Opinion Mining that is relevant to our work (COM). As proposed by Fang, Si, Somasundaram & Yu (2012), the aim of COM is to convey the opinions of various perspectives on a particular query topic and quantify their differences using an unsupervised topic model. COM is used to analyze debate records and breaking news stories. In contrast to keyword analysis, we compute different IR metrics from the content of news articles in order to evaluate and compare the bias in the SERPs of two search engines. Aktolga & Allan (2013) examine the sentiment toward controversial topics and proposes several diversification strategies based on the topic's sentiment. Their primary objective is to diversify a search engine's returned results based on various sentiment biases in blog articles, rather than to quantify bias in *news* search engine SERPs, as fulfilled in the first chapter of my thesis.

Demartini & Siersdorfer (2010) makes use of automatic and lexicon-based text classification techniques, Support Vector Machines and SentiWordNet, to extract sentiment value from the textual content of SERPs in response to controversial topics. Unlike, Demartini & Siersdorfer (2010) uses this sentiment information to compare the opinions expressed in the retrieved results of three commercial search engines without accounting for bias. In the first part of this thesis, a novel framework for evaluating bias in search engine results pages (SERPs) is presented that includes robust bias measures. Subsequently, in the second part, new bias methodology is proposed to measure gender bias in YouTube search results. Chelaru, Altingovde & Siersdorfer (2012) examine whether opinionated queries are sent to search engines by computing the sentiment of suggested queries for controversial topics. In a follow-up work (Chelaru, Altingovde, Siersdorfer & Nejdl, 2013), the authors employ a variety of classifiers to detect the sentiment conveyed in queries and expand upon prior studies with two new use cases. Instead of queries, throughout this work initially the SERPs in the news domain are analysed which requires to identify the stance of news articles, then YouTube video search results are examined to detect the gender of narrators.

2.2 Evaluating Fairness in Ranking

Fairness evaluation in ranked results has attracted attention in recent years. Yang & Stoyanovich (2017) propose three bias measures, namely Normalized discounted difference (rND), Normalized discounted Kullback-Leibler divergence (rKL) and Normalized discounted ratio (rRD), all of which are related to normalized discounted cumulative gain (NDCG) via the use of logarithmic discounting for regularization, as stated in the original paper. Researchers use these measures to determine if a group of individuals is subjected to systematic discrimination when only two distinct groups are included in a ranking: A protected (g_1) and an unprotected group (g_2). In other words, researchers quantify the relative representation of g_1 (the protected group), which consists of individuals who share a sensitive attribute such as race or gender. The definitions of these three proposed measures can be rewritten as follows:

$$f_{g_1}(r) = \frac{1}{Z} \sum_{i=10,20,\ldots}^{|r|} \frac{1}{\log_2 i} \left| d_{g_1}(i,r) \right|, \tag{2.1}$$

where $f(r)$ is a general definition of an evaluation measure for a given ranked list of documents, i.e. a SERP, and f_{g_1} is a definition particular to the protected group g_1. In this formulation, Z denotes a normalisation constant, r denotes the retrieved SERP's ranked list, and $|r|$ denotes the ranked list's size, i.e. the number of documents in the ranked list. Notably, i is intentionally increased by 10 to compute *set-based fairness* at discrete values such as *top*10, *top*20, and so on, rather than 1 as is typically done in IR to ensure that the proposed measures exhibit the correct behavior with larger sample sizes. The goal of computing the *set-based fairness* is to convey that being fair at higher positions on the ranked list, e.g. *top*10 vs. *top*100, is more crucial.

To quantify systematic bias, the rewritten formula establishes a distance function between the expected probability of retrieving a document belonging to g_1, i.e. in the overall population, and its observed probability at rank i. These probabilities turn out to be equal to P@n:

$$\mathrm{P}_{g_1}@n = \frac{1}{n} \sum_{i=1}^{n} [j(r_i) = g_1], \tag{2.2}$$

when computed over g_1 at the $|r|$ and i cut-off values for the three proposed metrics. In this formula, n is the number of documents in r that are treated as a cut-off value, and r_i is the document in r that is retrieved at rank i. Notably, $j(r_i)$ returns the label associated with the document r_i, stating whether it belongs to the g_1 or g_2 group. $[j(r_i) = g_1]$ denotes a conditional statement that returns 1 if the document r_i is a member of g_1 and 0 otherwise.

In the original paper, d_{g_1} is defined for rND, rKL, and rRD as:

$$
\begin{aligned}
d_{g_1}(i,r) =& \mathrm{P}_{g_1}@i - \mathrm{P}_{g_1}@|r| && \text{for rND,} \\
d_{g_1}(i,r) =& -\mathrm{P}_{g_1}@i \log\left(\frac{\mathrm{P}_{g_1}@|r|}{\mathrm{P}_{g_1}@i}\right) \\
& -(1-\mathrm{P}_{g_1}@i)\log\left(\frac{1-\mathrm{P}_{g_1}@|r|}{1-\mathrm{P}_{g_1}@i}\right) && \text{for rKL,} \\
d_{g_1}(i,r) =& \frac{\mathrm{P}_{g_1}@i}{1-\mathrm{P}_{g_1}@i} - \frac{\mathrm{P}_{g_1}@|r|}{1-\mathrm{P}_{g_1}@|r|} && \text{for rRD.}
\end{aligned}
$$

These measures, although inspired by IR evaluation measures, particularly in the context of content bias in search results suffer from the following limitations:

- The protected group (g_1) is the focus of the rND measure. If we compute f at steps of 1 with the desired proportion of the two groups equal to 50:50, the distance function of rND, indicated by d_{g_1}, would always return a value of 0.5 for the first retrieved document, where $i = 1$. This will always be the case, regardless of the group to which this text belongs, in our example, *pro* or *against*. This is due to the fact that d_{g_1} of rND is used in Eq. (2.1) with its absolute value. This is valid in our situation for $i = 1,2,4$ and $r = 10$, respectively, where we are measuring bias in the top-10 results. This is avoided in the original paper (Yang & Stoyanovich, 2017) by computing f at steps of 10, such as top-10, top-20, and so on, rather than at steps of 1, as is typically done in IR, which gives more meaningful results in our evaluation framework.
- The rKL measure is incapable of distinguishing between biases of same magnitude but in opposite directions when the desired proportion of the two groups is set to 50:50, i.e. bias toward *conservative* or *liberal* in our case. Additionally, in IR settings, the computed values from the KL-divergence (denoted as d_{g_1} for rKL) are more difficult to interpret than our measures, as our measures are based on common utility-based IR metrics. Additionally, because

KL-divergence generates larger distances for small datasets, it may compute larger bias values in the case of only 10 documents, and this situation may get even worse if we measure bias for a fewer number of documents, e.g. top-3, top-5 for a more fine-grained analysis. This disadvantage is avoided in the original study by computing the rKL values at discrete points of 10 instead of 1.

- The rRD measure does not treat protected and unprotected groups (g_1 and g_2) symmetrically, as indicated in the original paper, and hence does not apply to our framework. Our proposed measures treat g_1 and g_2 equally because we have two protected groups to measure bias in search settings: *pro* and *con* for stance bias, *conservative* and *liberal* for ideological bias. Furthermore, rRD is only applicable in special circumstances when g_1 is the minority group in the underlying population, as stated by the authors, but our measures do not face similar limitations in the context of search bias evaluation.

- These measures are concerned with differences in the relative representation of g_1 among distributions. Thus, from a broader perspective, additional samples are almost certainly required for these measures to exhibit predicted behavior and function properly. The original research conducts experiments using three distinct datasets: one synthetic dataset with 1000 samples and two real datasets with 1000 and 7000 samples, respectively, to evaluate bias using these measures, whereas we use only ten samples for query-wise evaluation. This is likely because these measures were developed primarily to quantify bias in ranked outputs rather than search engine results; none of these datasets contain search results either.

- These measures are difficult to apply in practice because they rely on a normalization term, Z, that is stochastically computed, i.e. as the maximum possible value of the corresponding bias measure for a given number of documents n and protected group size $|g_1|$. We utilize standard statistical tests in this study because they are easier to interpret, provide confidence intervals, and have previously been used successfully to examine disparities in search systems by Chen, Ma, Hannák & Wilson (2018).

- These measures ignore relevancy, a key aspect in determining search engine bias. For instance, in our case, when looking for a controversial topic, if the first retrieved document is about a news item belonging to g_1 but its content is irrelevant to the topic being searched, these measures will nevertheless deem the document positive for g_1. This document, on the other hand, has no effect on the user receiving an impartial representation of the controversial topic.

> This is because these measures were developed with the primary purpose of evaluating bias in ranked outputs rather than SERPs.

Zehlike, Bonchi, Castillo, Hajian, Megahed & Baeza-Yates (2017), based on Yang & Stoyanovich (2017)'s work, propose an algorithm to test the statistical significance of a fair ranking. Beutel, Chen, Doshi, Qian, Wei, Wu, Heldt, Zhao, Hong, Chi & others (2019) propose a pairwise fairness metric for recommender systems that adheres to the notion of equality of opportunity proposed by (Hardt, Price, Srebro & others, 2016). However, the authors, unlike us, base their assessment of fairness on personalized recommendations and ignore relevance, whereas the studies in this thesis are conducted in an unpersonalized IR setting that does not focus on a specific group, and we do consider relevance when measuring bias. Kallus & Zhou (2019) studies the cross-ROC curve and the corresponding xAUC metric for investigating the fairness of predictive risk scores as a bipartite ranking task, where the main goal is to rank positively labelled examples above negative ones via finding a good ranking function. However, their measures of bias based on the area under the ROC curve (AUC) are agnostic to the rank position at which a document was retrieved.

Although the proposed measures by Yang & Stoyanovich (2017) are beneficial for quantifying bias in ranked outputs in which individuals are ranked and some of these individuals are members of the protected group (g_1), they have the aforementioned drawbacks. These limitations are especially evident when evaluating content bias in a typical IR setting where web documents are ranked by search engines. In Chapter 3, these limitations are addressed by proposing a family of fairness-aware measures with the main purpose of evaluating content bias in SERPs in the context of stance and ideological search bias, based on standard utility-based IR evaluation measures.

In Chapter 6, the main focus is to measure *perceived* gender bias and for this purpose two new bias measures are proposed that compute percentage scores of bias which are easy to interpret to assess *equality of outcome* in YouTube video search results. Unlike the proposed measures in the scope of web search bias in Chapter 3 both of these new bias measures proposed in Chapter 6 have several advantages in the context of measuring *perceived* gender bias. These new bias measures take into account the YouTube video results only annotated with *male* and *female* gender labels meaning that the score of a male/female gender is computed over the sum of male and female scores. In this way, difference in computed metric values which shows the inequality between the gender groups becomes more significant and the scores of each gender group are symmetric around 0.5 (equal representation of male and female gender groups), which is the desired case. Additionally, the exposure measure with logarithmic weighting requires a normalisation term for interpretable results, yet the notion of an ideal list that was used by the researchers in Yang & Stoyanovich (2017) is not included since its definition differs with each ranked list in the dataset which is not practical to compute.

Geyik, Ambler & Kenthapadi (2019) present two measures in the scope of *equality of opportunity*, one of these measures is based on Yang & Stoyanovich (2017)'s measures and the other one compares the representation of a gender group in a given list with respect to the entire population. Lipani, Piroi & Yilmaz (2021) propose a measure which compares the representation of a categorical sensitive attribute in result documents with all the documents indexed by the search system. Gómez, Shui Zhang, Boratto, Salamó & Marras (2021) propose two measures which evaluate a given ranked list in terms of representation, i.e. proportion without including rank information, and exposure, i.e. incorporating stronger rank information with logarithmic weighting; both measures compare the given list with the full dataset. Thus, unlike the *equality of outcome* which is the group fairness criteria of this thesis, Geyik et al. (2019); Gómez et al. (2021); Lipani et al. (2021) propose measures to achieve *equality of opportunity.* Nonetheless, the presented measures are computed on search results by (Geyik et al., 2019; Lipani et al., 2021), while they are applied in recommendation settings by (Gómez et al., 2021). As discussed in Section 1.1, this thesis implements *equality of outcome* in search settings instead, for measuring search engine bias in news SERPs and gender bias in online educational materials.

2.3 Quantifying Bias

While the algorithms of search platforms are not transparent and not accessible to external researchers, algorithm auditing techniques enable systematic evaluation of results in a controlled environment (Sandvig, Hamilton, Karahalios & Langbort, 2014). Previous research has analyzed content bias using LDA-variant unsupervised methods and crowd-sourcing, as well as URL analysis for indexical bias.

Saez-Trumper, Castillo & Lalmas (2013) present unsupervised approaches for characterizing various types of bias in online news media and related social media communities, in addition to analyzing news sources' political views. Yigit-Sert, Altingovde & Ulusoy (2016) investigate media bias by analyzing user comments and the content of online news items in order to uncover latent facets of two extremely polarizing political issues in Turkey. Kulshrestha, Eslami, Messias, Zafar, Ghosh, Gummadi & Karahalios (2017) quantifies bias in social media by analyzing the author of a tweet, while in a follow-up work Kulshrestha et al. (2018) quantifies bias in web search by performing a URL analysis for Google in the political domain without conducting any SERP content analysis. The first part of this thesis focuses on content analysis to evaluate stance and ideological bias by examining Google and Bing SERPs from news sources such as the New York Times and BBC News. Then, the second part aims to measure gender bias through the narrator's *perceived* gender, i.e. from the viewer's perspective, in a given video.

Along with the unsupervised techniques, crowd-sourcing is a frequently utilized mechanism for detecting content bias. Crowd-sourcing is a commonly used method for labeling tasks in a variety of research fields, including image and video annotation (Krishna, Zhu, Groth, Johnson, Hata, Kravitz, Chen, Kalantidis, Li, Shamma & others, 2017; Vondrick, Patterson & Ramanan, 2013), object detection (Su, Deng & Fei-Fei, 2012), named entity recognition (Finin, Murnane, Karandikar, Keller, Martineau & Dredze, 2010; Lawson, Eustice, Perkowitz & Yetisgen-Yildiz, 2010), sentiment analysis (Räbiger, Gezici, Saygın & Spiliopoulou, 2018), and relevance evaluation (Alonso & Mizzaro, 2012; Alonso, Rose & Stewart, 2008). Yuen, King & Leung (2011) conduct an in-depth study of crowdsourcing applications. As Yuen et al. (2011) indicate, crowd-sourcing may also be used to elicit public opinions. Mellebeek, Benavent, Grivolla, Codina, Costa-Jussa & Banchs (2010) employ crowd-sourcing to classify Spanish consumer comments and demonstrates that non-expert Amazon Mechanical Turk (MTurk) annotations are a feasible and cost-effective alternative for expert annotations. In the first part of the thesis to measure stance and ideological bias, before attempting to establish an automated model for stance detection, crowd-sourcing is leveraged to collect public opinions on controversial topics rather than consumer products.

Apart from the content bias, another field of study is indexical bias. Indexical bias is a term that refers to bias that shows itself in the selection of items rather than the content of retrieved documents, i.e., content bias (Mowshowitz & Kawaguchi, 2002b). Researchers quantify merely indexical bias through the use of precision and recall measures in (Mowshowitz & Kawaguchi, 2002a). Additionally, the researchers approximate the *ideal* (i.e. norm) using the distribution generated by a collection of search engines for the purpose of measuring bias. However, this may not be a *fair* approach for evaluating bias, as the *ideal* should be objective, yet search engine results pages (SERPs) may contain *bias*. Similarly, Chen & Yang (2006) employ the same strategy to assess indexical and content bias; however, content analysis was carried out by representing the SERPs with a weighted vector containing various

HTML elements rather than performing an in-depth examination of the actual content. The first part of the thesis analyses the textual contents of Google and Bing SERPs to evaluate content bias rather than building the *ideal* using the SERPs of other search engines to ensure a more objective measurement of bias. On the other hand, the second part of the thesis only uses the clues about the narrator(s) in YouTube videos.

Along with the content categorization and indexical bias analysis, prior approaches for quantifying bias in auditing algorithms can be classified into three broad categories: *audience-based*, *content-based*, and *rater-based*. *Audience-based* measures are used to determine the political perspectives of media outlets and web pages by analyzing their users' interests, ideologies, and political affiliations, such as likes and shares on Facebook (Bakshy, Messing & Adamic, 2015), based on the premise that readers follow news sources that are ideologically similar to their own (Mullainathan & Shleifer, 2005). Lahoti, Garimella & Gionis (2018) model the ideological leaning of social media users and media sources on Twitter as a limited non-negative matrix factorisation problem. *Content-based* measures make use of linguistic elements contained in textual content; for example, Gentzkow & Shapiro (2010) extracts frequently used phrases from the Congress Reports of various political partisans (Democrats, Republicans). The researchers then devised the media slant index as a proxy for the political leanings of US newspapers. Finally, rater-based approaches make use of textual material and are thus comparable to content-based methods. In contrast to content-based methods, *rater-based* methods rely on subjective judgments about the sentiment, partisanship, or ideological leanings of content rather than linguistic analysis of the textual content. In general, approaches based on raters rely on crowd-sourcing to acquire labels for content analysis. For example, Budak, Goel & Rao (2016) quantify bias (partisanship) in US news outlets (newspapers and two political blogs) using 15 selected queries for a variety of controversial topics on which Democrats and Republicans disagree. The researchers use MTurk as a crowd-sourcing platform to determine the articles' content and political slant, i.e. whether they are favorable to Democrats or Republicans. Similarly, Epstein & Robertson (2017) use crowd-sourcing to score individual search results, and Diakopoulos, Trielli, Stark & Mussenden (2018) use the MTurk platform, a rater-based approach, to obtain labels for Google SERP websites by focusing on the content and utilizing prior work by Bakshy et al. (2015), an audience-based approach, specifically to quantify partisan bias. In this thesis, different approaches are leveraged to obtain annotations for quantifying bias. In the first part, Chapter 3, a rater-based technique utilizing the crowd-sourcing platform of MTurk is utilised to analyze web search bias via the stances and ideological leanings of news articles rather than partisan bias in the textual content of the SERPs. Similarly, Chapter 4 uses crowd-sourcing for measuring stance and ideological bias in search results with respect to location. Then, Chapter 5 attempts to follow a content-based approach using linguistic features to measure stance and ideological bias with an automated model. Then in the second part of the thesis, Chapter 6 leverages expert labels instead of crowd-sourcing for measuring *perceived* gender bias of narrators in YouTube video results. Lastly, Chapter 7 follows a content-based approach using audio features to track the source of *perceived* gender bias of narrators in YouTube.

There have been attempts to audit partisan bias on web search. Diakopoulos et al. (2018) present four case studies on Google search results, and in order to quantify partisan bias on the first page, they collect SERPs by issuing complete candidate names from the 2016 US presidential election as queries, and then using crowd-sourcing to obtain the SERPs' sentiment scores. They discovered that Google displayed a disproportionately high number of negative articles about Republican candidates compared to Democratic ones. Similarly, Epstein & Robertson (2017) present an election case study and employ a browser plugin to collect Google and Yahoo search data for election-related queries, then evaluate the SERPs using crowd-sourcing. Additionally, the researchers discovered a left-leaning bias, with Google being more biased than Yahoo. They discovered a modest but considerable ranking bias in the standard SERPs but not owing to personalisation (Robertson, Jiang, Joseph, Friedland, Lazer & Wilson, 2018) in their follow-up study. Similarly, researchers evaluate Google searches following Donald Trump's inauguration, utilizing a dynamic set of political queries and auto-complete suggestions (Robertson, Lazer & Wilson, 2018). Hu, Jiang, E. Robertson & Wilson (2019) conduct an algorithm audit and develop a specialized lexicon of partisan cues for determining the political partisanship of Google Search snippets in relation to their respective web pages. They describe the corresponding difference as bias for this particular use case without doing a robust user-perspective search bias review of SERPs.

Apart from partisan bias, recent studies have investigated *gender bias* in search. Chen & Yang (2006) examine gender bias in various resume search engines using a regression analysis in the context of individual and group fairness and found that there is a significant and consistent group unfairness against female candidates. Using a similar approach, Hannák et al. (2017) investigate *perceived gender* and *race* bias in two prominent online freelance websites to capture correlations between the profile features of workers and their reviews/ratings as well as their search rank position. They found that *perceived gender* and *race* bias are negatively correlated with search rank in one of these freelance websites. Kay et al. (2015) investigate the gender bias in image search results for a variety of occupations by only applying statistical significance tests without modelling the bias problem in the context of search and revealed that image search results exaggerate gender stereotypes and display the minority gender rather unprofessionally. Similarly, Singh et al. (2020) examine the image-based representation of highly gender-discriminated professions, e.g. nurse, computer programmer, in four digital platforms by simply comparing the ratio of male and female images in those platforms with the national labor statistics and found that women are largely underrepresented. Likewise, Otterbacher, Bates & Clough (2017) inspect gender stereotypes by directly computing the gender proportions in image search results returned by Bing without proposing any specific measures and showed that photos of women are more often retrieved for 'emotional' and similar traits, whereas men for 'rational' and related traits. In a follow-up work, by using a regression analysis the authors showed that sexist people are less likely to detect and report gender biases in search results (Otterbacher, Checco, Demartini & Clough, 2018). In addition to these studies which view gender bias in the context of search, some researchers also examine the relationship between how the course is displayed, i.e. the course is presented with a gender-inclusive photo, descriptions that contain more negative sentiment etc. which are the psychological cues, and the enrollment/engagement of different genders to STEM courses in online learning platforms (Brooks, Gardner & Chen, 2018; Kizilcec & Kambhampaty, 2020; Kizilcec & Saltarelli, 2019).

Raji & Buolamwini (2019) examine the impact of publicly naming biased performance results of commercial AI products in face recognition for directly challenging companies to change their products. Geyik et al. (2019) present a fairness-aware ranking framework to quantify bias with respect to protected attributes and improve the fairness for individuals without affecting the business metrics. The authors extended the measures proposed by Yang & Stoyanovich (2017), whose limitations discussed in Section 2.2, and evaluated their process using simulations applied to LinkedIn Talent Search. Vincent, Johnson, Sheehan & Hecht (2019) quantify search engines' dependency on user-generated content in order to respond to requests using Google search and Wikipedia pages. In another study, researchers offer a unique measure for assessing group fairness in sorted lists that incorporates users and their attention (Sapiezynski, Zeng, E Robertson, Mislove & Wilson, 2019). Gao & Shah (2019) present a methodology for estimating the solution space effectively and efficiently when fairness in information retrieval is modelled as an optimisation problem with a fairness constraint. The same researchers examine the top-k diversity fairness ranking in terms of statistical parity and disparate impact fairness and propose entropy-based metrics for quantifying the topical diversity bias in Google SERPs using clustering rather than a labelled dataset with group information.

Notable to remark here is the fact that, users typically pay more attention to top positions in a ranked list of search results which is called position bias and this phenomenon leads users to click those top positions with greater probability (Joachims, Granka, Pan, Hembrooke & Gay, 2005). Therefore, if search results are biased then users will be affected due to search engine manipulation effect (SEME) (Epstein & Robertson, 2015) and the impact is high if top positions are more biased. Since users tend to click the top positions with higher probability, this implicit user feedback will be logged, then fed to the ranking algorithm which will probably cause users to be exposed to an even higher bias – societal biases will be reinforced in the search results. Thus, even if the bias comes from the data itself, search platforms should still be responsible for mitigating it. As stated by Culpepper et al. (2018), an information retrieval system should be fair, accountable, and transparent. For preventing the severe consequences of bias in society through web search bias of stance and ideology of the web documents and gender bias in online learning, the first step is to reveal it, which is the main focus through this thesis, thereby further alerting users which could be effective in suppressing search engine manipulation effect (Epstein et al., 2017).

3. MEASURING WEB SEARCH BIAS

3.1 Introduction

As mentioned in Chapter 1, people are more susceptible to bias when they are not aware of it and alerting users could be a good remedy to alleviate SEME. Thus, the main aim of this chapter is to serve that purpose by presenting a new search bias evaluation framework in ranked lists to quantify bias in *news* SERPs. With the proposed robust fairness-aware IR measures, initially bias in the SERPs of two search engines is evaluated separately, then their relative bias is compared through incorporating relevance [1] and ranking information into the procedure without tracking the source of bias. The investigation of source of web search bias is fulfilled in Chapter 5.

The main contributions of this chapter can be summarised as follows:

- A *new generalisable search bias evaluation framework* is proposed to measure bias in SERPs by quantifying two different types of content bias which are stance and ideological bias.
- *Three novel fairness-aware measures of bias* are proposed which are based on common Information Retrieval (IR) *utility-based* evaluation measures: Precision at cut-off (P@n), Rank Biased Precision (RBP), and Discounted Cumulative Gain at cut-off (DCG@n) that do not suffer from the limitations of the previous bias measures as explained in Section 2.2 in detail.
- The proposed framework is applied to *evaluate the stance and ideological bias* of Google and Bing *news* search results for a variety of controversial topics,

[1] We are referring to the concept of relevance, which is defined in the literature as either system relevance or topical relevance, i.e. the relevance predicted by the system.

including but not limited to education, health, entertainment, religion, and politics.

- The framework is also used to *compare the relative bias* for queries related to various controversial issues on two widely used search engines: Google and Bing news search.

In the scope of this study, there are principally two reasons for opting for *equality of outcome*. To begin, not all corresponding debate questions (queries) pertaining to controversial topics have specific scientific answers. Second, identifying the stance for the entire ranking list, i.e. a set of documents that is a reasonably representative sample of the indexed texts, is too costly to crowd-source. Thus, the experiments become viable due to the adoption of an *ideal* ranking requiring *equal* representation.

It needs to be emphasised that stance and ideological leaning are differentiated in SERPs analysis. The stance in a SERP for a query topic can be pro or against, however the ideological leaning in a SERP denotes the specific ideological group, such as conservatives or liberals, that supports the relevant topic. Thus, a SERP's stance does not always imply the ideological leaning. For instance, given two controversial queries, "abortion" and "Cuba embargo", a SERP may take a positive stance on abortion, suggesting a liberal leaning, while taking a positive stance on Cuba embargo suggests a conservative leaning. Thus, only examining the SERPs' stance on controversial issues is insufficient and may possibly be misleading in assessing the ideological bias. It is worth noting that the conservative-liberal ideology space does not solely refer to political parties. In this context, the ideology labels are accepted as indicating a more conservative/liberal viewpoint on a specific controversial topic, as exemplified by Lahoti et al. (2018) for three popular controversial issues on Twitter: *gun control*, *abortion*, and *obamacare*.

Experiments demonstrate that there is no statistically significant difference in the amount of *stance bias* between the two search engines, indicating that none of them favours one particular stance over another. However, as proven by the "abortion" and "Cuba embargo" query examples, stance bias results should be interpreted warily. The polarisation of society is mostly driven by ideological leanings, and our second phase of experiments demonstrates a statistically significant difference in *ideological bias* between the two search engines, with one favouring one ideological leaning over the other.

In this section, first the search engine bias evaluation framework is described in more detail. Then, the measures of bias and the proposed protocol are presented to identify web search bias.

3.2.1 Preliminaries and Research Questions

Let $\mathcal{S}$ be the set of search engines and $\mathcal{Q}$ be the set of queries about controversial topics. When a query $q \in \mathcal{Q}$ is sent to a search engine $s \in \mathcal{S}$, the search engine s returns a SERP r. We define the stance of the i-th retrieved document r_i with respect to q as $j(r_i)$. A stance can have the following values: *pro*, *neutral*, *against*, *not-relevant*.

A document stance with respect to a topic can be:

- **pro** (👍) when the document is in favour of the controversial topic. The document describes more the pro aspects of the topic;
- **neutral** (✊) when the document does not support or help either side of the controversial topic. The document provides an impartial (fair) description about the pro and cons of the topic;
- **against** (👎) when the document is against the controversial topic. The document describes more the cons aspects of the topic;
- **not-relevant** (✖) when the document is not-relevant with respect to the controversial topic.

The analyses intentionally focus on recent *controversial* topics in the United States that are truly debatable, rather than on topics that may have been subject to false media balance, which occurs when the media portray opposing viewpoints as more equal than the evidence supports, e.g. the Flat Earth debate (Grimes, 2016; Stokes, 2019). The topic set includes abortion, illegal immigration, gay marriage, and similar *controversial issues* that encompass opposing viewpoints, as complicated concepts about one's identity, religion, political or ideological leanings are the actual areas where search engines are more likely to provide biased results and significantly influence people. The following research questions are addressed using controversial topics representing a broad range of issues in SERPs of Google and Bing through content analysis, i.e. analysing the textual content of the retrieved documents.

The first research question that this chapter aims to answer is:

RQ1: On a pro-against stance space, do search engines return *biased* SERPs towards controversial topics?

In order to answer RQ1, the degree of deviation of the ranked SERPs from an *ideal* distribution, i.e. an *equal* representation of different stances, is measured. To uncover bias caused by an imbalanced representation of diverse perspectives,the stances of documents are annotated via crowd sourcing and use these labels to evaluate stance bias.

The second research question is:

RQ2: Do search engines show *significantly different* magnitude of stance bias from each other towards controversial topics?

The second aim of this chapter is to identify bias in the distribution of ideological leanings represented in the SERPs' content. This is accomplished by linking each query $q \in \mathcal{Q}$ that pertains to a controversial topic with a particular ideological leaning. Then, the ideological bias of the content of a given SERP can be measured by combining the stances associated with each r_i and the associated ideological leaning of q. For example, if a topic is associated with a particular ideology and a document retrieved for this topic takes a pro stance, this document is considered to be biased towards that ideology.

We denote the ideological leaning of q as $j(q)$. The following values can represent an ideological leaning: *conservative*, *liberal*, *both or neither*.

A topic ideological leaning can be:

- **conservative** (🔴) when the topic is part of the conservative policies. The conservatives are in favour of the topic;
- **liberal** (🔵) when the topic is part of the liberal policies. The liberals are in favour of the topic;
- **both or neither** (○) when both or neither policies are either in favour or against the topic.

The last research question is:

RQ3: On a conservative-liberal ideology space, do search engines return *biased* SERPs and if so; are these biases *significantly different* from each other towards controversial topics?

RQ3 is naturally addressed by labelling each query topic as conservative or liberal, based on which ideology favours the claim in the query.Then the documents' stance

Table 3.1 Symbols, functions, and labels used throughout this chapter

Symbols	
$\mathcal{S}$	set of search engines.
s	a search engine $s \in \mathcal{S}$.
$\mathcal{Q}$	set of queries.
q	a query $q \in \mathcal{Q}$.
r	a ranked list of the given SERP (list of retrieved documents).
r_i	the document in r retrieved at rank i.
$\|r\|$	size of r (number of documents in the ranked list).
n	number of documents considered in r (cut-off).
Functions	
$j(r_i)$	returns the label associated to r_i.
$f(r)$	an evaluation measure for SERPs.
Labels	
👍	pro stance.
✊	neutral stance.
👎	against stance.
✖	not-relevant stance.
🔴	conservative ideological leaning.
🔵	liberal ideological leaning.
○	both or neither ideological leanings.

labels are interpreted within the conservative-to-liberal ideology[2] space and convert them to ideological leanings based on the corresponding query topics' assigned leaning labels. For instance, the topic of *abortion* relies on the query, *Should Abortion Be Legal?* Because the majority of liberals favour the statement in this query, *abortion* is assigned a liberal leaning. The stance labels of the documents returned in response to the query are converted to ideological leanings as follows. If a document takes a pro stance, which indicates that it supports the asserted idea, it has a liberal ideological leaning; if it takes an against stance, it has a conservative ideological leaning.

For reference, Table 3.1 shows a summary of all the symbols, functions and labels used in this chapter.

3.2.2 Measures of Bias

[2] We are referring to how crowd workers perceive ideology.

In the scope of this study, imbalanced representation of opinions refer to bias. Based on this, bias can be quantified by examining the degree to which the document distribution deviates from the *ideal*. Giving a broad definition of an ideal list brings problems; however, for the sake of this chapter, the presence of bias in a ranked list returned by a search engine is reported if the presented information *deviates significantly* from true likelihoods (White, 2013). As justified in Section 3.1, this study focuses on *equality of outcome*, which means that the true likelihoods of different perspectives are accepted as *equal* rather than computing them using the corresponding base rates. Thus, reversing the original definition, the *ideal* list becomes the one that minimises the difference between two opposing viewpoints, which can be referred to as 👍 and 👎 in the context of stances.

Formally, the *stance* bias in a SERP r is measured as follows:

$$\beta_f(r) = f_{\text{👍}}(r) - f_{\text{👎}}(r), \tag{3.1}$$

where f is a function that quantifies how likely it is that r would meet the user's information demand for the views 👍 and 👎. Note that ideological bias is also quantified in the same manner, by converting the documents' stances into ideological leanings, as discussed in Section 3.3.2. Prior to defining f, the mean bias (MB) of a search engine s is defined using Eq. (3.1) as follows:

$$\mathrm{MB}_f(s, \mathcal{Q}) = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \beta_f(s(q)).$$

A completely unbiased search engine would generate results with a mean bias of 0. A limitation of MB is that if a search engine is biased towards the 👍 perspective on one topic and towards the 👎 perspective on another, these two contributions will cancel out. To overcome this limitation, we define the mean absolute bias (MAB), which consists of calculating the absolute value of bias for each r. This is defined formally as follows:

$$\mathrm{MAB}_f(s, \mathcal{Q}) = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} |\beta_f(s(q))|. \tag{3.2}$$

A completely unbiased search engine would generate results with a mean absolute bias of 0. While the measure defined in Eq. (3.2) overcomes the problem of MB, MAB provides no information about towards which view the search engine is biased, making these two measures of bias complimentary.

In IR, the likelihood of r satisfying a user's information need is quantified using retrieval evaluation measures. Three utility-based evaluation measures were chosen from among these. This class of evaluation measures quantifies r in terms of its perceived value to the user and is often computed as the sum of the information gain multiplied by the number of relevant documents retrieved by r. P@n, RBP, and DCG@n are the three IR evaluation measures used in the following experiments.

P@n is formalised as in Eq. (2.2) for the *lps* viewpoint. Unlike the previous definition of $j(r_i)$, where the only potential outcomes were g_1 and g_2 for the document r_i, here j can return any of the labels associated with a stance (👍, ✊, 👎, and ✖). Thus, only pro and against documents that are relevant to the topic are considered, as $j(r_i)$ returns *neutral* and *not-relevant* otherwise. By substituting Eq. (2.2) for Eq. (3.1), the first measure of bias is obtained:

$$\beta_{\mathrm{P@}n}(r) = \frac{1}{n}\sum_{i=1}^{n}\Big([j(r_i) = \text{👍}] - [j(r_i) = \text{👎}]\Big).$$

The main shortcoming of this measure of bias is that it adopts a weak concept of ranking, in which the first n documents contribute equally to the bias score. By defining discount functions, the following two evaluation measures, RBP and DCG@n, avoid this issue.

RBP overcomes the aforementioned limitation in the following manner. Each document is weighted by RBP using the coefficients of a normalised geometric series with the value $p \in]0,1[$, where p is an RBP parameter. As with P@n, we reformulate RBP to assess bias as follows:

$$\mathrm{RBP}_{\text{👍}} = (1-p)\sum_{i=1} p^{i-1}[j(r_i) = \text{👍}]. \tag{3.3}$$

Substituting Eq. (3.3) for Eq. (3.1):

$$\beta_{\mathrm{RBP}}(r) = (1-p)\sum_{i=1} p^{i-1}\Big([j(r_i) = \text{👍}] - [j(r_i) = \text{👎}]\Big).$$

Instead, DCG@n utilises a logarithmic discount function to weight each document. As with P@n and RBP, DCG@n is formulated as follows to quantify bias:

$$\mathrm{DCG}_{\text{👍}}@n = \sum_{i=1}^{n} \frac{1}{\log(i+1)} [j(r_i) = \text{👍}]. \tag{3.4}$$

Substituting Eq. (3.4) for Eq. (3.1):

$$\beta_{\mathrm{DCG}@n}(r) = \sum_{i=1}^{n} \frac{1}{\log(i+1)} \Big([j(r_i) = \text{👍}] - [j(r_i) = \text{👎}] \Big) \tag{3.5}$$

Considering the web users, $n = 10$ is set for P@n and DCG@n, and $p = 0.8$ for RBP. Although this final formulation (Eq. (3.5)) resembles the rND measure, it does not suffer from the four shortcomings discussed in Section 2.2. Specifically, all of these presented bias measures: 1) do not concentrate on one group; 2) utilise a binary score associated with the document's stance or ideological leaning, similar to how these measures are applied in IR when incorporating relevance; similarly to IR 3) can be computed at each rank; 4) ignore non-relevant documents from the bias measurement and; the framework 5) provides a variety of user models for each of the three IR evaluation measures: P@n, DCG@n, and RBP.

3.2.3 Quantifying Bias

Using the bias measures introduced in Section 3.2.2, the bias of two search engines, Bing and Google, is quantified by analysing their news channels. Then, they are compared. Each step of the proposed protocol for quantifying SERP bias is described below.

- **News Articles in SERPs.**

  The controversial queries were obtained from ProCon.org (2018) and performed some filtering on the initial query set. After filtering, the size of the final query set was reduced to 57. A US proxy was used to submit each query in the final query set to the Google and Bing news search engines in the US. Then, the entire corpus returned by both engines in response to the queries in the set. Notably, the data collection method was carried out in a controlled environment to ensure that all requests were sent to the search engines concurrently. For more information on how the questions were chosen and the

SERPs were crawled, please refer to Section 3.3.1. After crawling and extracting the contents of all SERPs returned by both engines, the top-10 documents were annotated. Through crowd-sourcing, the stance of each document was labelled with respect to the queries and the ideological leanings of the queries were obtained. To identify the ideologies of documents, we converted their stance labels into ideological leanings based on the ideological orientation of their corresponding queries. Please refer to Section 3.3.1 for details about the crowd-sourcing campaigns as well as the transformation process.

- **Bias Evaluation.** The bias measures are computed for each SERPs using each of three IR-based bias measures: P@n, RBP, and DCG@n. The results are then aggregated using the two bias measures, MB and MAB.
- **Statistical Analysis.** To determine whether the bias found is not a result of randomness, a one sample t-test is applied: the null hypothesis is that there is no difference and that the true mean is equal to zero. If this hypothesis is rejected, it can be asserted that there is a statistically significant difference and the examined search engine is biased. Then, a two-tailed paired t-test is used to examine the difference in bias between the two search engines: the null hypothesis is that the difference between the two true means is equal to zero. If this hypothesis is rejected, then it can be asserted that there is a considerable difference in bias between the two search engines.

3.3 Experimental Setup

In this section, a description of the proposed experimental setup is provided based on the proposed method as defined in Section 3.2.3.

3.3.1 Dataset

All controversial issues were sourced from ProCon.org. ProCon.org is a non-charitable organisation dedicated to providing an online resource for conducting research on controversial issues. ProCon.org chooses themes that are controversial and significant to a large number of US residents, while also taking reader sug-

Table 3.2 All controversial topics, topics marked with red dots are conservative and blue for liberal

🔵 **Abortion**: Should Abortion Be Legal?	🔵 **Alternative Energy vs. Fossil Fuels**: Can Alternative Energy Effectively Replace Fossil Fuels?	🔵 **Animal Testing**: Should Animals Be Used for Scientific or Commercial Testing?
🔴 **Banned Books**: Should Parents or Other Adults Be Able to Ban Books from Schools and Libraries?	🔵 **Bill Clinton**: Was Bill Clinton a Good President?	🔵 **Born Gay? Origins of Sexual Orientation**: Is Sexual Orientation Determined at Birth?
○ **Cell Phones Radiation**: Is Cell Phone Radiation Safe?	🔵 **Climate Change**: Is Human Activity Primarily Responsible for Global Climate Change?	○ **College Education Worth It?**: Is a College Education Worth It?
🔴 **Concealed Handguns**: Should Adults Have the Right to Carry a Concealed Handgun?	🔴 **Corporal Punishment**: Should Corporal Punishment Be Used in K-12 Schools?	🔴 **Corporate Tax Rate & Jobs**: Does Lowering the Federal Corporate Income Tax Rate Create Jobs?
🔴 **Cuba Embargo**: Should the United States Maintain Its Embargo against Cuba?	○ **Daylight Savings Time**: Should the United States Keep Daylight Saving Time?	○ **Drinking Age - Lower It?**: Should the Drinking Age Be Lowered from 21 to a Younger Age?
🔴 **Drone Strikes Overseas**: Should the United States Continue Its Use of Drone Strikes Abroad?	○ **Drug Use in Sports**: Should Performance Enhancing Drugs (Such as Steroids) Be Accepted in Sports?	🔴 **Electoral College**: Should the United States Use the Electoral College in Presidential Elections?
🔵 **Euthanasia & Assisted Suicide**: Should Euthanasia or Physician-Assisted Suicide Be Legal?	○ **Vaping E-Cigarettes**: Is Vaping with E-Cigarettes Safe?	🔵 **Felon Voting**: Should Felons Who Have Completed Their Sentence (Incarceration, Probation, and Parole) Be Allowed to Vote?
○ **Fighting in Hockey**: Should Fighting Be Allowed in Hockey?	🔵 **Gay Marriage**: Should Gay Marriage Be Legal?	○ **Gold Standard**: Should the United States Return to a Gold Standard?
○ **Golf - Is It a Sport?**: Is Golf a Sport?	🔵 **Illegal Immigration**: Should the Government Allow Immigrants Who Are Here Illegally to Become US Citizens?	🔵 **Israeli-Palestinian Two-State Solution**: Is a Two-State Solution (Israel and Palestine) an Acceptable Solution to the Israeli-Palestinian Conflict?
○ **Lowering the Voting Age to 16**: Should the Voting Age Be Lowered to 16?	🔵 **Medical Marijuana**: Should Marijuana Be a Medical Option?	○ **Milk - Is It Healthy?**: Is Drinking Milk Healthy for Humans?
🔵 **Minimum Wage**: Should the Federal Minimum Wage Be Increased?	🔵 **National Anthem Protest**: Is Refusing to Stand for the National Anthem an Appropriate Form of Protest?	🔵 **Net Neutrality**: Should Net Neutrality Be Restored?
🔵 **Obamacare**: Obamacare Is the Patient Protection and Affordable Care Act (Obamacare) Good for America?	🔵 **Obesity a Disease?**: Is Obesity a Disease?	○ **Olympics**: Are the Olympic Games an Overall Benefit for Their Host Countries and Cities?
○ **Penny - Keep It?**: Should the Penny Stay in Circulation?	○ **Police Body Cameras**: Should Police Officers Wear Body Cameras?	🔴 **Prescription Drug Ads**: Should Prescription Drugs Be Advertised Directly to Consumers?
🔵 **Prostitution - Legalize It?**: Should Prostitution Be Legal?	🔵 **Right to Health Care**: Should All Americans Have the Right (Be Entitled) to Health Care?	🔴 **Ronald Reagan**: Was Ronald Reagan a Good President?
🔵 **Sanctuary Cities**: Should Sanctuary Cities Receive Federal Funding?	🔴 **School Uniforms**: Should Students Have to Wear School Uniforms?	🔵 **School Vouchers**: Are School Vouchers a Good Idea?
○ **Social Media**: Are Social Networking Sites Good for Our Society?	🔴 **Social Security Privatization**: Should Social Security Be Privatized?	🔴 **Standardized Tests**: Is the Use of Standardized Tests Improving Education in America?
🔵 **Student Loan Debt**: Should Student Loan Debt Be Easier to Discharge in Bankruptcy?	○ **Tablets vs. Textbooks**: Should Tablets Replace Textbooks in K-12 Schools?	🔵 **Teacher Tenure**: Should Teachers Get Tenure?
🔴 **Under God in the Pledge**: Should the Words "Under God" Be in the US Pledge of Allegiance?	🔵 **Universal Basic Income**: Is Universal Basic Income a Good Idea?	○ **Vaccines for Kids**: Should Any Vaccines Be Required for Children?
🔵 **Vegetarianism**: Should People Become Vegetarian?	○ **Video Games and Violence**: Do Violent Video Games Contribute to Youth Violence?	○ **Voting Machines**: Do Electronic Voting Machines Improve the Voting Process?

gestions into account. All 74 controversial issues were extracted from the website, along with their associated topic questions. Then, for practical reasons and without pre-selecting any themes, three filters were applied to these topics. The first filter restricts the analysis to *polar questions*, commonly known as yes-no questions, due to their lack of opposing sides. This filter reduced the size of the topic set from 74 to 70. The second filter eliminates topics that do not have up-to-date material on their ProCon.org topic pages, as they are not recent controversial issues and so would not provide up-to-date results. The second filter reduced the number of topics to 64. Finally, the third filter includes topics only if both search engines yield results for the matching topic questions which are queries; otherwise, comparative analysis is impossible. After the final filter, the size of the final topic set decreased from 64 to 57. Table 3.2 contains the full list of controversial topic titles with questions used in this study.

These 57 topics were crawled using the topic questions. For instance, the topic title 'abortion' contains the question 'Should Abortion Be Legal?'. The topic questions reflect the primary debates around the related controversial issues, which were used exactly as-is (i.e. with upper-case letters and without eliminating punctuation) while querying the search engines.

To avoid any personalisation effect, the news search results were collected in *incognito mode*. Thus, the retrieved SERPs are not tailored to any particular user, but (presumably) to all US users. Each topic question, query, was submitted to Google and Bing's news search engines using a US proxy. Because the news versions of the two search engines were used, any sponsored results that might have influenced the analysis did not appear in the news search results. Then, the URLs of the retrieved results were crawled for the same topic question (query) in order to minimise time lags between search engines, as the SERP for a given query may change over time. Following that, the crawled URLs were used to extract the textual contents of the top-10 documents. Thus, the time gap between Google and Bing's SERPs for each controversial issue (entire corpus) was reduced to an average of 2-3 minutes. Additionally, prior to initiating the crawling process, some experiments were conducted in the news search and default search using a small set of topics (different from the topic set provided in Table 3.2) and no significant changes were observed, particularly in the top-10 documents of the news search, even with 10-15 minute time lags. This suggests that the news search engine is less dynamic than the default search engine channel, and it has been estimated that the 2-3 minute time lags would have little effect on the search results.

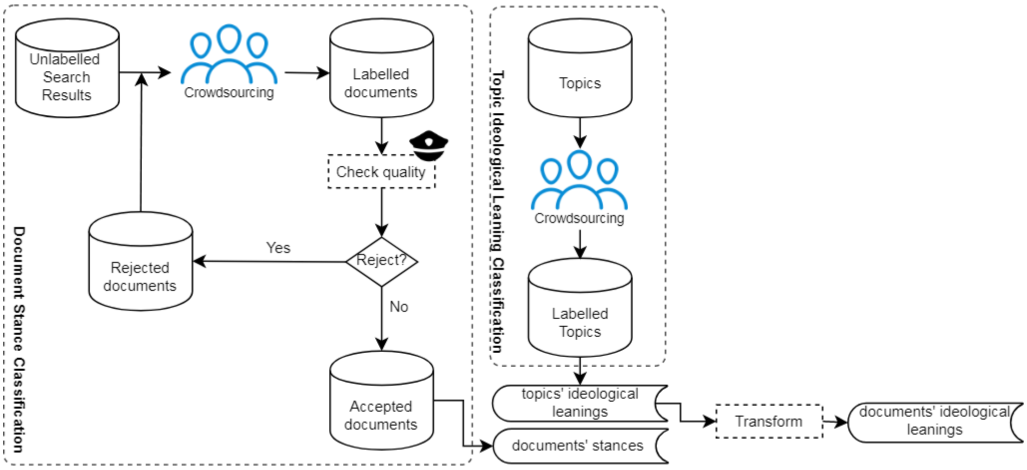

Document Stance Classification
Unlabelled Search Results
Crowdsourcing
Labelled documents
Check quality
Reject?
Yes
No
Rejected documents
Accepted documents
Topic Ideological Leaning Classification
Topics
Crowdsourcing
Labelled Topics
topics' ideological leanings
documents' stances
Transform
documents' ideological leanings

Figure 3.1 Flow-chart of the crowd-sourcing campaigns

3.3.2 Crowd-sourcing Campaigns

The flowchart in Figure 3.1 illustrates the end-to-end process for acquiring stances and ideological leanings. The flowchart's highlighted (dotted) components depict the Document Stance Classification (DSC) and Topic Ideological Leaning Classification (TILC) processes.

The DSC process takes unlabeled top-10 search results that have been crawled using the data collection procedure described in Section 3.3.1 and crowd-sources the stance labels for all of these documents in relation to the topic questions ($\mathcal{Q}$) that were used to retrieve them. As illustrated in the flowchart, the TILC process leverages crowd-sourcing to identify the ideological leanings of all topic questions ($\mathcal{Q}$). Then, all the stance labels that have accepted through the DSC process, translated into ideological leaning labels depending on the ideology provided to their respective topic questions. The process of collecting document labels for stance and ideological leaning detection is detailed below.

To label the stance of each document with respect to the topic questions ($\mathcal{Q}$) crowd-sourcing was used and Mturk was chosen as a platform. To ensure high-quality crowd-labeling, the following task properties were specified in this platform. Due to the majority of the issues being relevant to the US, crowd workers from the US were hired for the annotation tasks. Additionally, skilled and experienced professionals were attempted to be recruited by establishing the following criteria: The approval rate for Human Intelligence Tasks (HITs), i.e. single, self-contained task for a worker, should be better than 95% and the number of HITs approved should be greater than 1000 for each worker. Per each HIT, the wage was set as 0.15$ and a time limit was 30 minutes. Three crowd-workers judged each document.

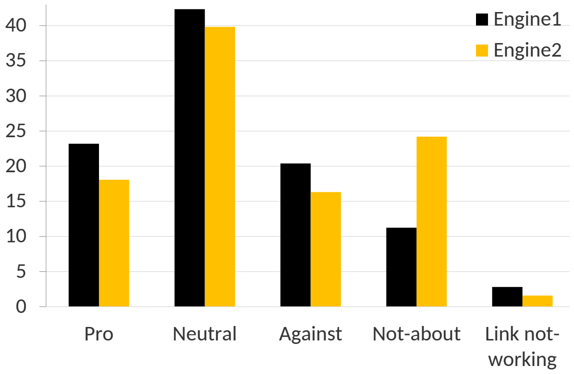


Figure 3.2 Percentages of the document stance labels annotated by crowd-workers

To classify a document's stance, crowd-workers were asked to label it as pro, neutral, against, irrelevant, or link not-working after being presented with a controversial topic question. Prior to assigning a task, a worker received instructions in three categories, ranging from general to specific. To begin, workers were given an overview of the stance detection task, followed by a list of the task's steps, such as reading the topic question, opening the news article link, and so on, followed by guidelines and suggestions. The final part of an HIT defined the stance labels *pro*, *neutral*, and *against* as defined in Section 3.2.1. Additionally, there was a hint for workers, stating that while the title of the article may give a rough indication of the article's stance, it may not be sufficient to determine the article's overall opinion. Then, workers were requested to read the remainder of the article as well. Apart from these, a disclaimer was included at the bottom of the page informing workers that some of the responses were already known and that their HITs, or single, self-contained tasks for workers, may be rejected based on evaluation. Then, on the following page, the worker was presented with an HIT that included a topic question (query), a link to a news article whose stance will be determined by repeating/reminding the the stance detection task's main question.

To achieve reliable annotations, first a randomly selected group of documents were annotated, which were then used to assess the quality of crowd-labels, as mentioned in the warning to workers. Using these expert labels, the low quality annotations were rejected and requested new labels for those documents. This method was repeated until all of the document labels were received. At the conclusion of this iterative process, two agreement scores were computed on the accepted labels for document stance detection, which are listed in Table 3.3. Inter-rater agreement scores are expressed as a percentage agreement between corresponding annotators. The pairwise agreement was applied, entering a value of 1 if there is agreement and 0 otherwise. Following that, the mean for the fractions was calculated. The reported Kappa score for classification of document stances is deemed to be in *fair* agreement. Previously, researchers reported a Kappa score of the inter-rater agreement between experts (0.385) rather than crowd-workers for the same task, i.e. document stance classification in SERPs towards a different query set that includes controversial topics as well as popular products, claiming MTurk workers struggled with the task (Alam & Downey, 2014). Although the stance detection task in this study appears to be more difficult, as the queries are limited to controversial topics, the reported Kappa score for MTurk workers is comparable to their expert agreement score. This indicates that the annotator agreement level is sufficient given the subjective nature and difficulty of the task.

Figure 3.2 shows the distribution of accepted stance labels for each search engine's search results. One could argue that when a news search engine receives a query about a controversial topic, the SERP will predominantly contain controversial articles that support one prevailing viewpoint on the subject. As a result, informational websites or articles adequately covering various perspectives on the subject, i.e. documents with a neutral stance, would never be included in the study. However, as illustrated in Figure 3.2, this reasoning is refuted by the fact that the majority of labels for both search engines are actually *neutral*.

To identify each topic's ideological leaning, crowd-sourcing was once more used, as illustrated in Figure 3.1. Crowd-workers were requested to label each topic as *conservative*, *liberal*, *both, or neither.* To ensure that high-quality annotations were obtained for topic ideology detection as well, worker attributes were configured to be identical to those used for stance detection. Likewise, crowd-workers were recruited solely from the US. The wage per HIT was set to 0.1$ and the time limit was set to five minutes. Similarly to the stance detection task, an overview was presented, steps were described, and the informational page was concluded with guidelines and tips. For this task, the final part included the ideological leaning definitions from Section 3.2.1. Additionally, workers were asked to assess the ideological leaning of a given topic in light of the current ideological climate and warned about the possibility of their HITs being rejected. On the next page, workers were presented an HIT with a topic question (query), i.e. one of the topic's main debates, and asked the following: *Which ideological group would respond favourably to this query?* The topics attributed to conservative or liberal leanings were determined by a majority vote of five annotators. The topics' leanings are summarised in Table 3.2. Table 3.3 additionally includes two agreement scores computed on the judgments for detecting ideological leanings.

To map the documents' stances from *pro-to-against* to *conservative-to-liberal*, a simple transformation was applied. This modification is necessary for documents that

Table 3.3 Agreement among Crowd-workers

Campaign	Inter-rater	Fleiss-Kappa
Document Stance	0.4968	0.3500
Topic Ideological Leaning	0.5281	0.3478

Table 3.4 The search engines' performance, as determined by the p-values of a two-tailed paired t-test performed on engines 1 and 2.

	P@10	RBP	DCG@10
Engine 1	0.8509	0.7708	3.9114
Engine 2	0.7404	0.6886	3.4773
p-value	< 0.001	< 0.001	< 0.01

take a pro-abortion or pro-Cuba embargo stance. While both documents take the same stance, they have opposite ideological leanings, as a pro-abortion stance implies a liberal leaning, whereas a pro-Cuban embargo stance implies a conservative leaning. For some topics (such as the Cuba embargo), the *pro-to-against* stance labels are interpreted as the *conservative-to-liberal* ideological leaning labels in search results, whilst for others (such as abortion), we as the *liberal-to-conservative.* On the other hand, for topics such as *vaccines for children*, where the crowded-label resulted in *both or neither*, the *conservative-to-liberal* or *liberal-to-conservative* conversion was insignificant and hence removed from the analysis. Note that due to budget restrictions, the crowd-sourcing protocol was developed to generate high-quality crowd-labels by labelling (expert) a random sample of documents, using an iterative process, and deciding on these labels by majority vote.

3.3.3 Results

In Table 3.4 the performance of the two search engines is presented. This is evaluated across all topics. When a document is classed as pro, con, or neutral, it is considered relevant. The difference is statistically significant for all evaluation measures.

In Table 3.5 stance bias of the search engines is demonstrated. Note, for all three bias measures, P@10, RBP, and DCG@10, a lower value indicates less bias in the context of this work as opposed to their respective classical IR measures. For all three IR evaluation measures, all MB and MAB scores are positive. Additionally, the differences in MB and MAB measures between the two search engines are statistically not significant, as demonstrated by the two-tailed pair t-test on these measures. The

Table 3.5 The search engines' stance bias, as evaluated by the p-values of a two-tailed paired t-test performed on engine 1 and engine 2.

		P@10	RBP	DCG@10
MB	Engine 1	0.0281	0.0197	0.1069
	Engine 2	0.0175	0.0271	0.1142
	p-value	> 0.05	> 0.05	> 0.05
MAB	Engine 1	0.2596	0.2738	1.3380
	Engine 2	0.2246	0.2266	1.0789
	p-value	> 0.05	> 0.05	> 0.05

Table 3.6 The search engines' ideological bias, as evaluated by the p-values of a two-tailed paired t-test performed on engine 1 and engine 2.

		P@10	RBP	DCG@10
MB	Engine 1	-0.1368	-0.1247	-0.6290
	Engine 2	-0.1289	-0.1386	-0.6591
	p-value	> 0.05	> 0.05	> 0.05
MAB	Engine 1	0.2579	0.2894	1.3989
	Engine 2	0.2184	0.2158	1.0456
	p-value	> 0.05	< 0.05	< 0.05

ideological bias is shown in Table 3.6. As with Table 3.5, the lower the value, the better, since the same bias measures are being used. Table 3.6 is analogous to Table 3.5. In contrast to Table 3.5, all MB scores are negative for all three IR evaluation measures, whereas all MAB scores are positive. The two-tailed paired t-test performed on MBs to compare the bias difference between engines 1 and 2 is statistically not significant. Nevertheless, the two-tailed paired t-test on MABs is statistically not significant for the measure P@10, but is statistically siginificant for the measures RBP and DCG@10.

In Figure 3.3, the distribution of topic-specific SERPs over the pro-against stance space for the DCG@10 measure is depicted. The x-axis represents the pro-stance score ($DCG_{👍}$@10), whereas the y-axis represents the against-stance score ($DCG_{👎}$@10). Each point represents a topic's overall SERP score. The black points represent the SERPs that engine 1 retrieved, whereas the yellow points represent the SERPs that engine 2 retrieved. The overall stance bias score ($\beta_{DCG@10}$) of SERPs for each topic evaluated on the two search engines is visualized in Figure 3.5. The x-axis represents engine 1, while the y-axis represents engine 2. Positive coordinates denote topics whose SERPs are skewed towards the pro stance, whereas negative coordinates denote topics whose SERPs are skewed towards the against stance.

Figures 3.4 and 3.6 are similar to Figures 3.3 and 3.5, except that they measure ideological bias in the former case rather than stance bias. As a result, Figure 3.4

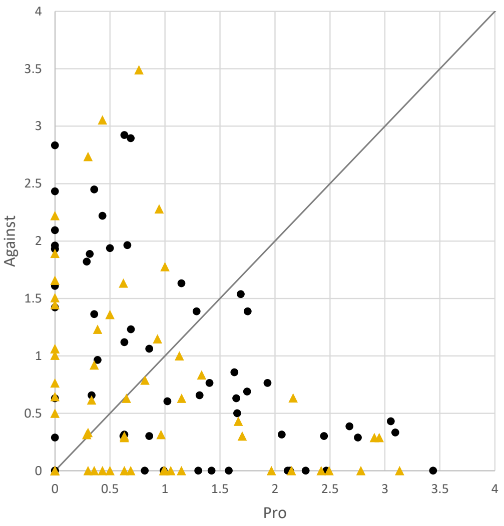


Figure 3.3 *DCG*👍@10 against *DCG*👎@10 measured on stances – black points for engine 1 and yellow points for engine 2

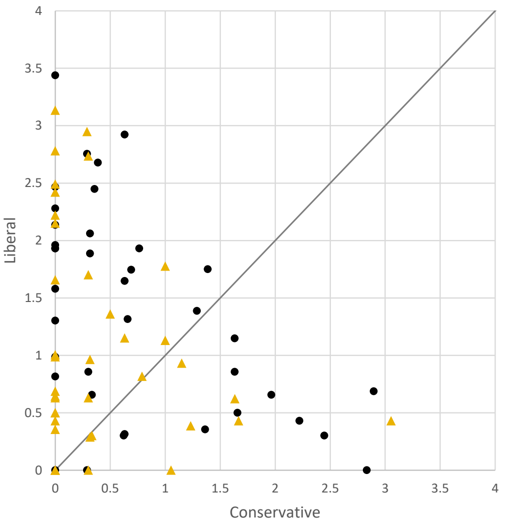


Figure 3.4 *DCG*🔴@10 against *DCG*🔵@10 measured on ideological leanings – black points for engine 1 and yellow points for engine 2

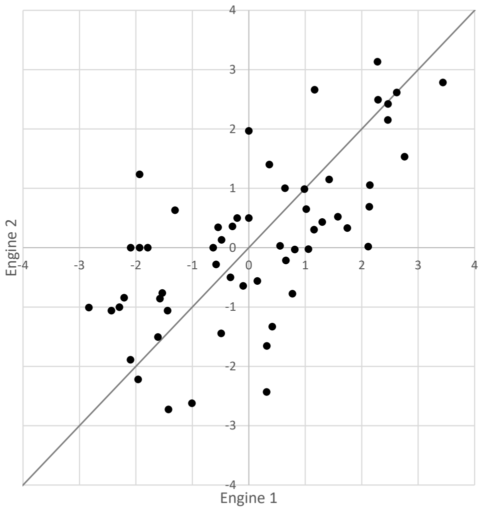


Figure 3.5 $\beta_{DCG@10}$ measured on stances, where positive is 👍 and negative is 👎

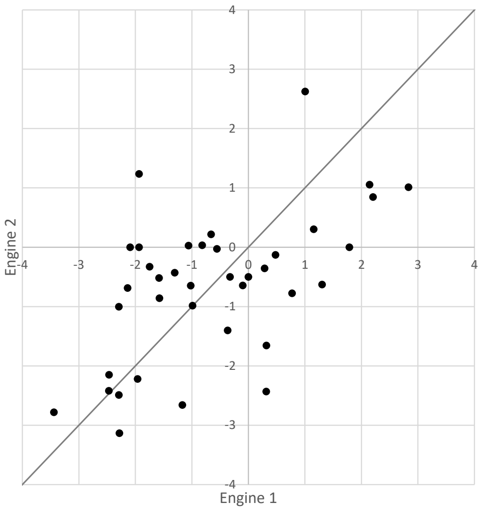


Figure 3.6 $\beta_{DCG@10}$ measured on leanings, where positive is 🔴 and negative is 🔵

illustrates the distribution of overall SERPs for topics throughout the conservative-liberal ideological spectrum using the DCG@10 measure. Similarly, in Figure 3.6, the overall ideological bias score ($\beta_{DCG@10}$) is compared, that is, the difference between conservative and liberal leaning scores, of the SERPs, where positive coordinates represent topics with a conservative leaning and negative coordinates represent topics with a liberal leaning.

3.4 Concluding Discussion

Prior to examining the potential bias in SERPs, the retrieval performances of two different search engines are compared. As seen in Table 3.4 both search engines work effectively, but engine 1 outperforms engine 2 by a statistically significant margin. This is corroborated by the three IR evaluation measures.

Following that, it is determined whether the search engines return biased results in terms of document stances (RQ1) and, if so, then it is further examined whether the engines show the same level of bias (RQ2), indicating that the difference between the engines is not statistically significant. In Table 3.5, all MB scores are positive, and the engines appear to be biased in favour of the pro stance with regard to RQ1. The one-sample t-test is used to determine the presence of stance bias, that is, whether the true mean is different from zero, as discussed in Section 3.2.3. However, because these biases are statistically not significant, this expectation may be due to noise — there is no systematic stance bias, i.e. preference for one stance over another. On the basis of MAB scores, it is clear that both engines exhibit an absolute bias. The two-tailed t-test, however, demonstrates that the difference between the two engines is not statistically significant. These findings demonstrate that neither search engine is biased in favour of a *particular* stance when providing results, as there is no statistically significant deviation from the *ideal* distribution. Nonetheless, both engines have an absolute bias that can be viewed as the expected bias for a topic question. These empirical findings imply that search engines are biased in pro stance for some topics and in against stance for others.

In Figure 3.3, the results are shown. The values in this figure are those used to calculate the MAB score for the DCG@10 column. It illustrates that the difference between both engines' pro and against stances on topics is equally distributed. To emphasise, no topic can be located in the plot's upper-right area because the sum of their coordinates is limited by the highest DCG@10 score. Additionally, it is observed that the engines agree in the majority of cases. This is also verified in Figure 3.5, where the stance bias scores ($\beta_{DCG@10}$), that is, the difference between the DCG@10 scores for the pro stance and the DCG@10 scores for the against stance, of topics are balanced between the upper-right and lower-left quadrants. Additionally, these two quadrants represent the area of agreement between the two engines in terms of stance. The remaining two quadrants are for topics on which the engines disagree. It can be stated that, in the majority of cases, the engines agree.

Lastly, it is determined whether search engines are ideologically biased (RQ3). By examining the MB scores in Table 3.6, it is clear that both search engines appear to be ideologically aligned in the same direction — liberal (all MB scores are negative). Unlike the stance bias, a one-sample t-test on MB scores indicates that these expectations are statistically significant with different confidence levels, i.e. p-value 0.005 across all three IR measures for engine 2; p-value 0.05 on RBP and DCG@10 for engine 1. These findings suggest that both search engines have a liberal bias. When the two search engines' MB scores are compared, it is revealed that their differences are statistically not significant, implying that the observed difference could be due to random noise. Due to the fact that all MAB scores are positive, it is possible to observe that both engines exhibit an absolute bias. In contrast to the stance bias, there is a difference in expected ideological bias between the two search engines this time. For RBP and DCG@10, there is a statistically significant difference between the engines. This conclusion, together with the diverse user models modelled by various evaluation measures, suggests that users' perceived bias may vary in response to their behaviour. A user who always inspects the top-10 results (as modelled by P@10) may view engine 1 and engine 2 to have the same ideological bias, whereas a less systematic user who only inspects the top results may consider engine 1 to be more biased. Additionally, when this conclusion is compared to the engines' performance, it is clear that the engine with the better performance is more biased than the engine with the worse performance.

When comparing Figures 3.4 and 3.3, it is clear that the points in Figure 3.4 are less evenly distributed than those in Figure 3.3. The majority of the topics discussed are liberal. Additionally, engine 2 is more conservative than engine 1 in terms of points. When comparing Figures 3.6 and 3.5, it is clear that the engines in Figure 3.6 are more biased towards the liberal than the engines in Figure 3.5. Additionally, the engines agree on the majority of the points — the majority of the points are located in the upper-right and lower-left quadrants.

In this chapter, new bias evaluation measures and a generalisable evaluation framework were introduced to address the issue of web search bias in news search results. The proposed framework was utilised to quantify stance and ideological bias in Bing and Google SERPs and further to compare their relative bias towards controversial topics. The initial results show that both search engines seem to be unbiased when considering the document stances and *ideologically* biased when considering the document ideological leanings. The main intention of this chapter is to analyse SERPs without the effect of personalisation. Thus, these results highlight that search biases exist even though the personalization effect is minimized and that search engines can empower users by being more accountable.

To conclude, this chapter focuses on stance and ideological bias merely in top-10 SERPs without including any additional information, e.g. location or personalisation, in the analysis. Since designing a controlled bias study to measure stance/ideological bias in personalised SERPs seems to be quite complicated, Chapter 4 will investigate the effect of localisation in unpersonalised search settings. It is critical to emphasise that identifying the source of bias is not the purpose of this chapter, therefore the results can only be interpreted as a potential indicator. Chapter 5 will make multiple attempts to trace the source of bias by utilising automatic stance detection methods rather than crowd-sourcing to get document labels, and thus analysing bias over the entire corpus of retrieved SERPs. Nonetheless, the problem is viewed from the user's perspective, and regardless of the source of the bias, the results are biased in the way described. The findings appear to corroborate prior research (Diakopoulos et al., 2018; Epstein & Robertson, 2017) suggesting that liberal (left-leaning) partisan bias exists in SERPs; even in unpersonalised search settings (Robertson et al., 2018).

4. THE IMPACT OF GEOLOCATION ON SEARCH BIAS

4.1 Introduction

In Chapter 3, web search bias has been analysed for Google and Bing in the US. In this chapter, the impact of location on web search bias is investigated as well. For this, online search bias is analysed for the UK and US versions of Bing and Google. Specifically, this chapter aims to answer the following research questions.

The first research question is:

RQ1: On a conservative-liberal ideology space, do search engines return *biased* SERPs and if so; are these biases *significantly different* from each other towards controversial topics?

In order to answer RQ1, similar to the Section 3.2.1, the degree of deviation of the ranked SERPs from an *equal* representation of different ideologies, is measured. Further, the magnitude of ideological bias in Google and Bing is compared.

The second research question is:

RQ2: On a conservative-liberal ideology space, do different geolocations affect the existence of *ideological* bias in search engines?

Specifically in the scope of this chapter, the effect of location is examined. Initially, the existence of bias is analysed in the UK version of Bing (Google) and the US version of Bing (Google) to check if different locations affect the existence of ideological bias in each search engine.

The last research question is:

RQ3: On a conservative-liberal ideology space, do different geolocations affect the magnitude of *ideological* bias difference in search engines?

In addition to investigating the effect of location on the existence of bias, it is examined whether different locations influence the magnitude of bias. For this, the bias of the UK version of Bing (Google) and the US version of Bing (Google) is compared in terms of level of bias they show.

4.2 Experimental Setup

In this section, the steps of the evaluation procedure to examine the location effect on web search bias are described. These evaluation steps are similar to the ones described in Section 3.2.3.

4.2.1 Dataset

For the dataset crawling, the subset of the controversial queries obtained from ProCon.org (2018) in Chapter 3, were used. In the scope of this chapter, among the query list displayed in Table 3.2, only the queries of the controversial topics that has ideological leanings were used for the analysis of location. In Chapter 3, results showed that neither of the US search engines are biased in terms of stance but ideology and they are biased towards the liberal leaning. Thus, to investigate the impact of location, only the queries that could be used in ideological bias were selected for the bias analysis. For this purpose, all the 38 topics with their queries (topic questions) marked with red and blue dots in Table 3.2 were leveraged to crawl the top-10 SERPs of Bing and Google using the locations of the UK and US.

For crawling the SERPs, the news channels of UK and US Google and Bing search engines were used. To avoid any personalisation effect, news search results were collected in *incognito mode*. Thus, the retrieved SERPs are not tailored to any particular user, but (presumably) to all UK/US users. Each topic question, query, was submitted to UK/US Google and Bing's news search engines using UK/US proxies. Note that the news channel do not show any sponsored results in the SERPs. First, the URLs of the retrieved SERPs were crawled for the same topic question (query) in order to minimise time lags between search engines in the same location and specifically for this chapter in different locations as well. Thus, news SERPs of Bing and Google were automatically crawled using the UK and US proxies in parallel. In this way, the time gap between Bing-US and Google-US and their UK counterparts was attempted to be minimised. Since it has been known that news channel is less dynamic than the default web channel, the minimised time lags would probably little effect on the search results. Following that, the crawled URLs were used to extract the textual contents of the top-10 documents.

4.2.2 Crowd-sourcing Campaigns

For the annotation of the dataset, a similar procedure depicted in Section 3.3.2 was fulfilled. To label the stance of each document with respect to the topic questions (queries), crowd-sourcing was used and MTurk was chosen as a platform. To ensure high-quality crowd-labeling, the following task properties were specified in this platform. Although the majority of the issues being relevant to the US, in the scope of this work for the annotation of UK SERPs, crowd workers from the UK and for the US SERPs crowd workers from the US were hired for the annotation. Additionally, skilled and experienced professionals were attempted to be recruited by establishing the following criteria: The approval rate for Human Intelligence Tasks (HITs), i.e. single, self-contained task for a worker, should be better than 95%. Per each HIT, the wage was set as 0.02$, i.e. in the scope of web search bias each HIT contained 5 annotations instead of 1 annotation, and a time limit was 30 minutes. Three crowd-workers judged each document. Although a similar crowd sourcing procedure was attempted to apply, lower inter-rater agreement scores were obtained as 0.3215 and 0.2979 for the UK and US SERPs annotations respectively. This is probably because, a detailed iterative process could not be fulfilled due to time and budget constraints. To map the documents' stances from *pro-to-against* to *conservative-to-liberal*, a simple transformation was applied. For the details, please refer to Section 3.3.2.

4.2.3 Quantifying Bias

Using the bias measures introduced in Section 3.2.2, the bias of four search engines, the UK and US versions of Bing and Google, is quantified by analysing their news channels. In the scope of this chapter, the proposed protocol in Section 3.2.3 has been applied to measure bias by investigating the impact of location as well.

While the one-sample t-test is applied to check the existence of bias in search engines separately, the two-tailed paired t-test is used to check whether the magnitude of bias difference between two search engines is statistically significant. Unlike the protocol as described in Section 3.2.3, in the scope of location analysis, Bonferroni correction (Sedgwick, 2012) has also been applied since there are many hypotheses to be checked using t-tests. Bonferroni correction is used for multiple hypothesis testing and in the context of this bias analysis, there are 36 hypotheses in total. Hence, without the Bonferroni correction, with the significance level, $\alpha = .05$ and 36 hypotheses, the probability of identifying at least one significant result due to chance is around 0.84. Note that for the significance level where $\alpha = .05$, and with the Bonferroni correction new $\alpha = .00138$ and Bonferroni correction rejects the null hypothesis for each p-value (p_i) if $p_i <= .00138$ instead of .05. For the significance level where $\alpha = .01$, new $\alpha = .00028$ and for $\alpha = .001$, new $\alpha = .000028$ and so on.

4.2.4 Results

The impact of location has been investigated mainly in two ways, first on overall ideological bias results, then on ideological bias of each search engine separately.

The bias measures proposed in Section 3.2.2 for measuring the stance/ideological bias in Chapter 3 were computed on the new location-based dataset. Prior to examining the existence of bias, first Google and Bing's retrieval performances were measured for the UK and US locations independently.

In Table 4.1 and Table 4.2 it is observed that Bing and Google show similar retrieval performances – the two-tailed paired t-tests computed on retrieval scores are statistically not significant for the UK and US locations. This is verified across all three IR evaluation measures. Nonetheless, it is observed that the US versions of Bing and Google show higher retrieval performances than the UK. For this, in Table 4.3 and Table 4.4, the retrieval performances of engine 1 and engine 2 were assessed for the UK and US locations. The results show that the US versions of both search engines show higher retrieval performances than their UK counterparts. The two-tailed paired t-tests computed on the retrieval scores are statistically significant for engine 1 and engine 2 in Table 4.3 and Table 4.4 respectively. This is verified across all three IR evaluation measures.

Following that, it is determined whether Google and Bing return biased results in the UK and US locations. In Table 4.5, both UK search engines are ideologically biased towards conservative (all MB scores are positive) – one sample t-tests computed on MB and MAB scores are statistically significant. The bias scores for the ideological leanings of conservative and liberal bias scores cancelled each other out, thus MBs show lower scores than MABs. Similarly, both US search engines seems to be ideologically biased in Table 4.6, however one sample t-tests computed on MB scores are statistically not significant but computed on MAB scores are statistically significant. This is probably because, the bias scores for the ideological leanings of conservative and liberal bias scores cancelled each other out. Unlike the UK search engines, neither of the US engines are biased. In addition, there is no difference in the magnitude of bias – two-tailed paired t-tests computed on MBs and MABs are statistically not significant for both the UK and US search engines. This is verified across all three IR evaluation measures.

Apart from these, it has also been investigated if the same search engine (engine 1 or engine 2) shows similar level of bias in different locations. In Table 4.7, the UK version of engine 1 seems to be more biased towards conservative than its US counterpart. Yet, two-tailed paired t-tests computed on MB scores are statistically not significant with Bonferroni correction for the measures P@10, RBP, and DCG@10. In terms of MABs, both the UK and US versions of engine 1 show similar level of absolute bias – two-tailed paired t-tests computed on MAB scores are statistically not significant and this is confirmed by all three IR evaluation measures. For engine 2, in Table 4.8, two-tailed paired t-tests computed on MB scores are statistically not significant for P@10, while statistically significant for RBP, and DCG@10 due to Bonferroni correction ($p - values$ = .0113, .0013, and .0006 for P@10, RBP, and DCG@10 respectively). Similar to engine 1, both the UK and US versions of engine 2 also show similar level of absolute bias – two-tailed paired t-tests computed on MAB scores are statistically not significant and this is confirmed by all three IR evaluation measures.

In Figure 4.1, the distribution of query-specific SERPs over the conservative-liberal ideological spectrum for the DCG@10 measure in the UK is depicted. The x-axis represents the conservative-ideological score ($DCG_{\color{red}\bullet}$@10), whereas the y-axis represents the liberal-ideological score ($DCG_{\color{blue}\bullet}$@10). Each point represents a query's overall SERP score. The black points represent the SERPs that engine 1 retrieved, whereas the yellow points represent the SERPs that engine 2 retrieved. Similarly, in Figure 4.2, the distribution of topic-specific SERPs over the conservative-liberal ideological spectrum in the US is displayed. The overall ideological bias score ($\beta_{DCG@10}$) of SERPs for each query evaluated on the two UK search engines is visualized in Figure 4.3. The x-axis represents engine 1, while the y-axis represents engine 2.

Table 4.1 The UK search engines' performance, as determined by the p-values of a two-tailed paired t-test performed on engines 1 and 2.

	P@10	RBP	DCG@10
Engine 1	0.6027	0.5896	2.9203
Engine 2	0.6649	0.6170	3.0993
p-value	> 0.05	> 0.05	> 0.05

Table 4.2 The US search engines' performance, as determined by the p-values of a two-tailed paired t-test performed on engines 1 and 2.

	P@10	RBP	DCG@10
Engine 1	0.9730	0.8734	4.4305
Engine 2	0.9838	0.8790	4.4691
p-value	> 0.05	> 0.05	> 0.05

Table 4.3 The location-wise performance of engine 1, as determined by the p-values of a two-tailed paired t-test performed on the UK engine 1 and US engine 1.

	P@10	RBP	DCG@10
Engine 1 (UK)	**0.6027**	**0.5896**	**2.9203**
Engine 1 (US)	**0.9730**	**0.8734**	**4.4305**
p-value	$<$ **0.0001**	$<$ **0.0001**	$<$ **0.0001**

Table 4.4 The location-wise performance of engine 2, as determined by the p-values of a two-tailed paired t-test performed on the UK engine 2 and US engine 2.

	P@10	RBP	DCG@10
Engine 2 (UK)	**0.6649**	**0.6170**	**3.0990**
Engine 2 (US)	**0.9838**	**0.8790**	**4.4691**
p-value	$<$ **0.0001**	$<$ **0.0001**	$<$ **0.0001**

Positive coordinates denote topics whose SERPs are skewed towards the conservative leaning, whereas negative coordinates denote topics whose SERPs are skewed towards the liberal leaning. Similar to Figure 4.3, Figure 4.2 displays the overall ideological bias score ($\beta_{DCG@10}$) of SERPs for each query evaluated on the US search engines.

4.2.5 Concluding Discussion

Before evaluating the possibility of bias in SERPs, the retrieval performances of two distinct search engines are compared. As seen in Table 4.1 and Table 4.2 both search

Table 4.5 The UK search engines' ideological bias, as evaluated by the p-values of a two-tailed paired t-test performed on engine 1 and engine 2.

		P@10	RBP	DCG@10
MB	Engine 1	0.1108	0.1214	0.5740
	Engine 2	0.1027	0.1339	0.6260
	p-value	> 0.05	> 0.05	> 0.05
MAB	Engine 1	0.1378	0.1573	0.7205
	Engine 2	0.1622	0.1873	0.8829
	p-value	> 0.05	> 0.05	> 0.05

Table 4.6 The US search engines' ideological bias, as evaluated by the p-values of a two-tailed paired t-test performed on engine 1 and engine 2.

		P@10	RBP	DCG@10
MB	Engine 1	-0.0027	0.0107	-0.0065
	Engine 2	-0.0405	-0.0693	-0.2718
	p-value	> 0.05	> 0.05	> 0.05
MAB	Engine 1	0.1811	0.2039	0.9247
	Engine 2	0.1865	0.2354	1.0623
	p-value	> 0.05	> 0.05	> 0.05

Table 4.7 The location-wise ideological bias of engine 1, as evaluated by the p-values of a two-tailed paired t-test performed on the UK engine 1 and US engine 1.

		P@10	RBP	DCG@10
MB	Engine 1 (UK)	0.1108	0.1214	0.5740
	Engine 1 (US)	-0.0027	0.0107	-0.0065
	p-value	> 0.05	> 0.05	> 0.05
MAB	Engine 1 (UK)	0.1378	0.1573	0.7205
	Engine 1 (US)	0.1811	0.2039	0.9247
	p-value	> 0.05	> 0.05	> 0.05

Table 4.8 The location-wise ideological bias of engine 2, as evaluated by the p-values of a two-tailed paired t-test performed on the UK engine 2 and US engine 2.

		P@10	RBP	DCG@10
MB	Engine 2 (UK)	0.1027	**0.1339**	**0.6260**
	Engine 2 (US)	-0.0405	**-0.0693**	**-0.2718**
	p-value	**< 0.05**	**< 0.05**	**< 0.05**
MAB	Engine 2 (UK)	0.1622	0.1873	0.8829
	Engine 2 (US)	0.1865	0.2345	1.0623
	p-value	> 0.05	> 0.05	> 0.05

engines work effectively in the UK and US respectively, but engine 2 outperforms engine 1 yet the difference is statistically not significant. This is verified by the three IR evaluation measures. Also, it is observed that the US versions of both search

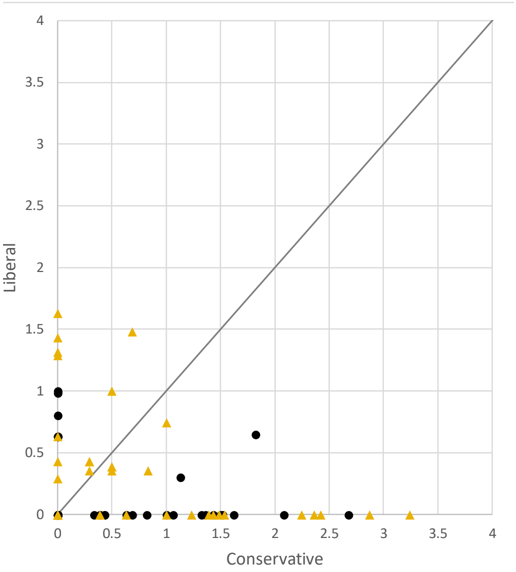


Figure 4.1 DCG🔴@10 against DCG🔵@10 measured on ideological leanings of UK – black points for engine 1 and yellow points for engine 2

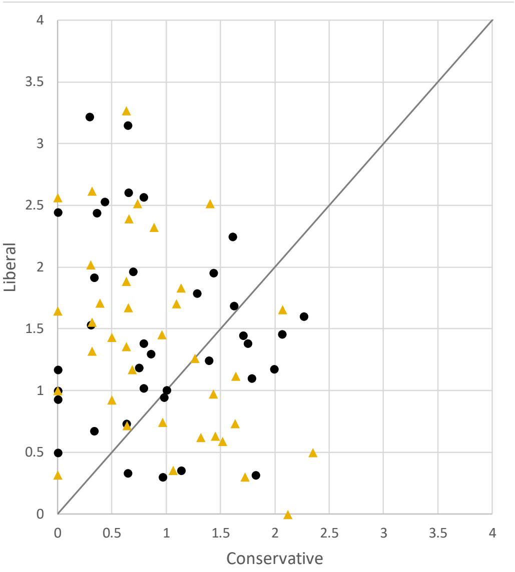


Figure 4.2 DCG🔴@10 against DCG🔵@10 measured on ideological leanings of US – black points for engine 1 and yellow points for engine 2

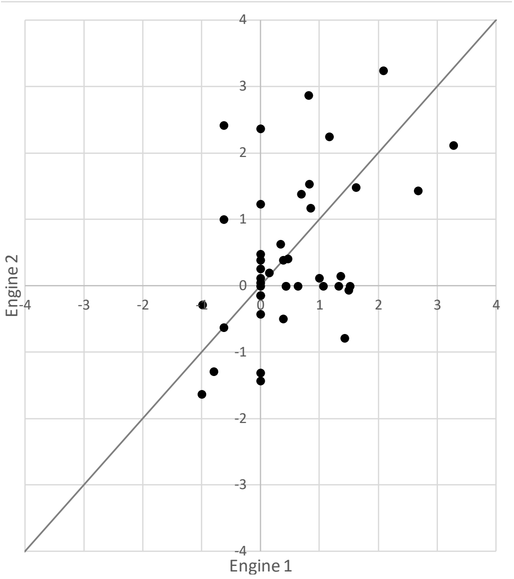


Figure 4.3 $\beta_{DCG@10}$ measured on UK leanings, where positive is 🔴 and negative is 🔵

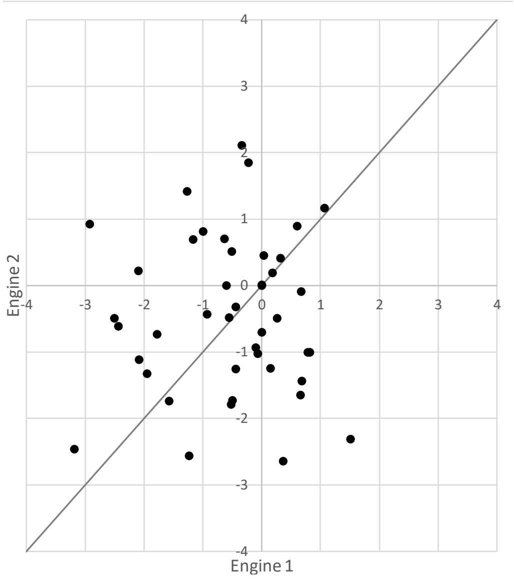


Figure 4.4 $\beta_{DCG@10}$ measured on US leanings, where positive is 🔴 and negative is 🔵

engines work better than their UK counterparts. For this, the retrieval of the same search engine (Google or Bing) is compared in the UK and US locations to check whether the location affects the retrieval performance, or not. In Table 4.3, in terms of retrieval performance the US version of engine 1 outperforms its UK counterpart

– the two tailed paired t-test provides statistically significant results and it is verified across the three IR evaluation measures. Likewise, in Table 4.4 the US version of engine 2 outperforms its UK counterpart as well.

Then, it is determined whether search engines return biased results in terms of ideology leanings (**RQ1**) and, if this is the case, it is then determined if the search engines exhibit the same level of bias (**RQ1**), suggesting the difference between engines is statistically not significant. In Table 4.5, all MB scores are positive, and the UK search engines appear to be biased in favour of the conservative leaning with regard to **RQ1**. The one-sample t-test is used to determine the presence of stance bias, that is, whether the true mean is different from zero, as discussed in Section 4.2.3. The one-sample t-test computed on MB scores is statistically significant for both search engines. Nonetheless, in order to answer the **RQ1**, two-tailed paired t-test computed on MBs are statistically significant which means that the UK versions of the search engines show similar level of bias. On the basis of MAB scores, it is clear that both engines exhibit an absolute bias. The two-tailed t-test demonstrates that the difference between the two engines is statistically not significant. This means that both search engines show similar level of absolute bias. Unlike, in Table 4.6, all MB scores are negative meaning that the US versions of the search engines seem to be biased towards the liberal learning with respect to **RQ1**. Yet, the one-sample t-test computed on MB scores is statistically not significant so neither of the US search engines are biased. The one-sample t-test computed on MAB scores, it is observed that both engines exhibit an absolute bias. The two-tailed paired t-test computed on MAB scores is statistically not significant, thus both search engines exhibit similar level of bias (**RQ1**). Regarding the **RQ2**, the UK versions of both search engines are biased towards conservative, while neither of the US search engines are biased.

Regarding the **RQ3**, in Table 4.7, the two-tailed paired t-test computed on MBs and MABs is statistically not significant which means that location does not affect the magnitude of bias that the engine 1 exhibits. Unlike, in Table 4.8, the two-tailed paired t-test computed on MBs are statistically significant for RBP, and DCG@10 meaning that location affects the magnitude of bias in the case of the engine 2. Based on MAB scores, the engine 2 exhibits the same level of bias irrespective of the location. In Figure 4.1, both engine 1 and engine 2 seems to be biased towards the conservative leaning – the query points are generally appear to be far away from the trendline. Unlike, in Figure 4.2, the query points are more dispersed and most of them are close to the trendline – neither of the search engines seem to be clearly biased. In Figure 4.3, both search engines appear to be biased towards conservative, whereas in Figure 4.4, the query points are more dispersed and there is no visible bias towards a specific ideological leaning. These interpretations are consistent with our aforementioned conclusions inferred from the results.

5. INVESTIGATING THE SOURCE OF SEARCH BIAS

5.1 Introduction

As discussed in Section 3.1, bias can occur as a result of either the input data, which may contain biases, or the search algorithm, which includes sophisticated features and carefully chosen algorithms that, while designed to be effective at satisfying information needs, may introduce systematic biases. Thus, in order to make a compelling case that search engines are biased, one must investigate the source of the bias and demonstrate that the bias is inherent in the search algorithm. Yet, throughout the thesis the source of bias has not been investigated. Chapter 3 focused on stance/ideological bias in Google and Bing SERPs in unpersonalised search settings without any localisation and the source of bias analysis. Then, Chapter 4 only incorporated location information into the bias analysis without tracing the source of bias. Hence, this chapter mainly provides several attempts to investigate the source of bias. This chapter mainly aims to answer to the following research question:

RQ: What is **the source of web search bias (if exists)**, does it come from the input data, or the ranking algorithm?

For this purpose, this chapter presents different state-of-the-art approaches and further customise them to automate the annotation procedure. Since crowd-sourcing is a costly process to obtain labels for the whole corpus, source of bias analysis requires an automated model for obtaining labels as the initial step.

Table 5.1 Best Evaluation Results on the Stance Dataset with the fine-tuned BERT Model

Model	Pro	Against	Neutral	Not-rel
Fine-tuned BERT Model	0.65	0.38	0.42	0.89

5.2 State-of-the-art Approaches

In the scope of this chapter, several state-of-the-art approaches have been leveraged to get annotations automatically. Nonetheless, the automated model is expected to give sufficiently good class-wise F1-scores in order to be used for the annotation since the annotation accuracy will a high impact on bias results. First of all, the pre-trained BERT (Devlin, Chang, Lee & Toutanova, 2018) model was fine-tuned on the crowd-labelled stance dataset of 839 documents in the context of Chapter 3. Then, to improve the model results, more documents were manually annotated with expert labels. At the end, the total number of annotated documents has become 3573 including 829 pro, 693 against, 1516 neutral, 535 not-relevant instances. Note that for the model-training phase, more irrelevant instances were generated to make the dataset more balanced. Nonetheless, generally there are more neutral documents in the crawled SERPs as displayed in Figure 3.2. The best evaluation results by fine-tuning BERT model on the new dataset are displayed in Table 5.1. Since the class-wise F1-scores are not sufficiently good, the following state-of-the-art approaches were applied as well to improve the automated model capability.

Table 5.2 F1-scores for SVM, Random Forest, and XGBoost

	Pro	Against	Neutral	Not-rel
SVM	0.50	0.19	0.55	0.59
Random Forest	0.53	0.10	0.46	0.43
XGBoost	0.49	0.34	0.61	0.84

5.2.1 Traditional Machine Learning Models

Since transformer-based models are prone to overfitting especially in the case of fine-tuning on small datasets, with the initial annotated small set of 839 documents (annotated in the context of Chapter 3), several traditional machine learning models were used to overcome the over-fitting problem. Support Vector Machine (SVM),

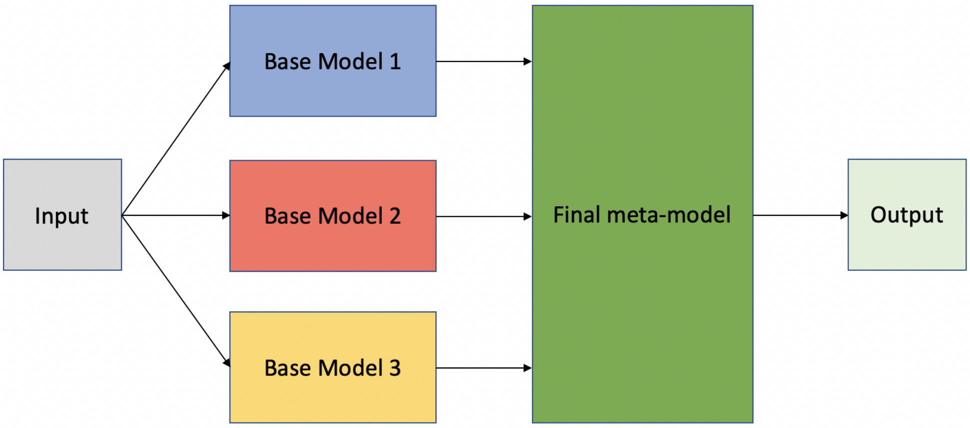


Figure 5.1 Model Stacking[1]

Random Forest, and XGBoost were applied and among these models, XGBoost gave the best class-wise F1-scores as displayed in Table 5.2. Yet, this model is still not sufficient to be used for annotation in bias analysis.

5.2.2 Model Stacking

Further, model stacking was applied with the aim of creating a strong classifier using the weak classifiers as depicted in Figure 5.1. This approach slightly improved the results, yet not sufficient.

5.2.3 Universal Language Model Fine-tuning (ULMFiT)

The steps of the proposed Universal Language Model Fine-tuning (ULMFit) (Howard & Ruder, 2018) pipeline which focuses on transfer learning for domain-adaptation, were applied on the annotated dataset. The ULMFiT is visualised in Figure 5.2. For applying ULMFit, initially the pre-trained BERT model was fine-tuned on the SERPs dataset without stance labels, this step is called as language model fine-tuning. In this intermediate step, different methods for the fine-tuning proposed in the original paper, namely slanted triangular learning rates, discriminative fine-tuning, and gradual unfreezing were applied as well. Lastly, the fine-tuned

[1] `https://towardsdatascience.com/a-practical-guide-to-stacking-using-scikit-learn-91e8d021863d`

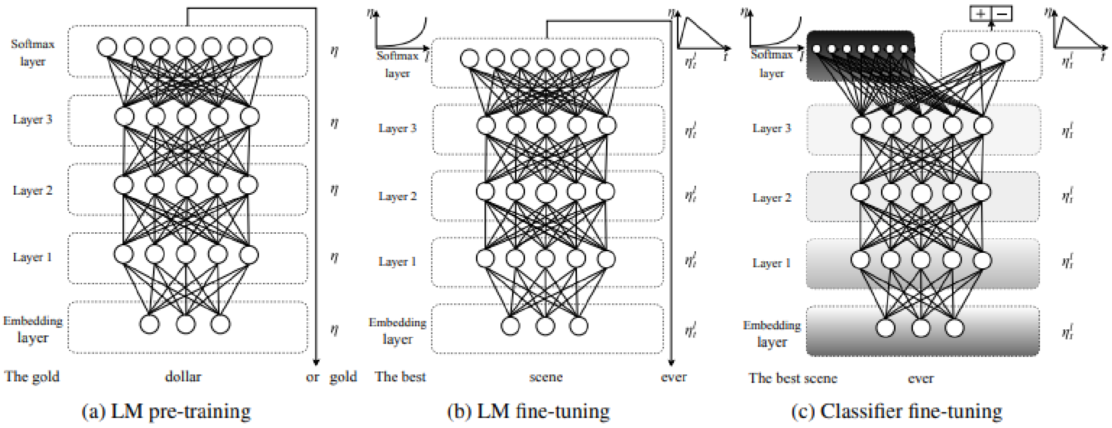


Figure 5.2 ULMFiT (Howard & Ruder, 2018)

language model was fine-tuned again on the annotated dataset with stance labels for stance classification task. For the visualisation of the ULMFiT, please s This pipeline helps to improve the classification performance, especially on small datasets. However, the results did not show a noticeable improvement.

5.2.4 Longformer: The Long-Document Transformer

Since the popular transformer-based models such as BERT (Devlin et al., 2018), a more robust version of BERT namely RoBERTa (Liu, Ott, Goyal, Du, Joshi, Chen, Levy, Lewis, Zettlemoyer & Stoyanov, 2019), and a lite-version of BERT namely ALBERT (Lan, Chen, Goodman, Gimpel, Sharma & Soricut, 2019) are unable to process long sequences due to their self-attention mechanisms, researchers proposed a new transformer called as Longformer (Beltagy, Peters & Cohan, 2020) to process long sequences. The lack of ability to process long sequences might cause poor learning since there are many long web document contents in our dataset. Hence, Longformer was evaluated for the annotation task as well. Yet, the results did not show a big improvement.

5.2.5 Stable Fine-tuning & Mixout

Apart from experimenting with different state-of-the-art approaches, lastly hyperparameter tuning was also applied to improve the automated model results. Despite

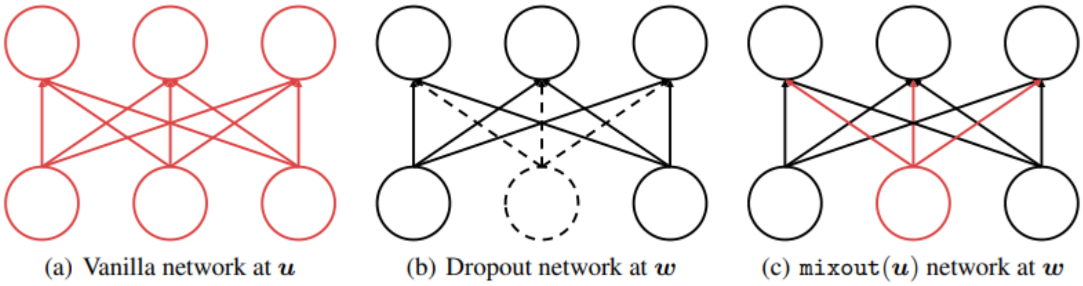

(a) Vanilla network at $\boldsymbol{u}$
(b) Dropout network at $\boldsymbol{w}$
(c) mixout($\boldsymbol{u}$) network at $\boldsymbol{w}$

Figure 5.3 The regularisation technique of mixout motivated by dropout (Lee et al., 2019)

the strong empirical performance of fine-tuned transformer models, fine-tuning is known to be an unstable process, different random seeds can result in large variance of the task performance (Mosbach, Andriushchenko & Klakow, 2020). Thus, Mosbach et al. (2020) focus on this stability issue and report the best hyper-parameter values for BERT, RoBERTa and ALBERT to alleviate the fine-tuning instability. Since the fine-tuning instability has also been observed in the context of the experiments with transformer-based models in this chapter, the recommended hyper-parameter values were leveraged to overcome the instability issue. Although experimenting with the recommended hyperparameters helped to achieve a more stable fine-tuning process regardless of the random seed, i.e. similar results were obtained with different training runs, the automated model could not achieve sufficiently good class-wise F1-scores.

Table 5.3 Best Evaluation Results on the Stance Dataset with the Aforementioned State-of-the-art Approches

Model	Pro	Against	Neutral	Not-rel
The Best Result	0.58	0.76	0.52	0.93

In the scope of hyperparameter tuning and fine-tuning instability, lastly the technique of mixout was applied to regularise the fine-tuning of a pre-trained model motivated by another widely-used regularisation technique of dropout Srivastava, Hinton, Krizhevsky, Sutskever & Salakhutdinov (2014), especially for small datasets as displayed in Figure 5.3. It has been observed that although BERT-large outperforms BERT-base generally [2], fine-tuning might fail if the target training dataset has a relatively small size, i.e. generally less than 10,000 instances. Similar to the recommended hyperparameters, mixout also helped to achieve a more stable fine-tuning process for transformed-based models especially. Although, a significant improvement was observed in fine-tuning stability, the automated models did not

[2] https://huggingface.co/transformers/v3.3.1/pretrained_models.html

give sufficiently good results that can be used for obtaining stance labels. The best results that have been achieved using the aforementioned approaches can be seen in Table 5.3.

As you can see that in comparison to our previous model results, we achieved better results – especially the decrease in loss value is quite high but our class-wise accuracies are still not good enough for annotation in bias analysis.

5.3 Concluding Discussion

Based on the results in Table 5.3, it has been observed that the results of the various aforementioned state-of-the-art approaches to establish an automated for obtaining stance labels did not give sufficient class-wise F1-scores. Since crowd-sourcing would be a very costly process to obtain labels for the whole corpus without an automated model, in the scope of this thesis, the source of search engine bias has not been investigated.

6. MEASURING GENDER BIAS IN ONLINE EDUCATION

6.1 Introduction

Students are increasingly using online materials to learn new subjects or to supplement their learning process in educational institutions. Issues regarding gender bias have been raised in the context of formal education and some measures have been proposed to mitigate them. However, online educational materials in terms of possible gender bias and stereotypes which may appear in different forms are yet to be investigated in the context of search bias in a widely-used search platform. As a first step towards measuring possible gender bias in online platforms, YouTube educational videos have been investigated in terms of the perceived gender of their narrators. Bias measures for ranked search results to evaluate educational videos returned by YouTube in response to queries related to STEM (Science, Technology, Engineering, and Mathematics) and NON-STEM fields of education. Gender is a research area by itself in social sciences which is beyond the scope of this chapter. In this respect, for annotating the perceived gender of the narrator of an instructional video only a crude classification of gender into male, and female is used. Then, for analysing perceived gender bias, bias measures that have been inspired by search platforms are utilised and further rank information is incorporated into our analysis. The preliminary results demonstrate that there is a significant bias towards the male gender on the returned YouTube educational videos, and the degree of bias varies when we compare STEM and NON-STEM queries. Finally, there is a strong evidence that rank information might affect the results.

YouTube states that they audit their machine learning systems to avoid cases leading to gender discrimination (YouTube, 2018). However, this does not guarantee that the returned videos are not biased towards a specific gender. In this gender bias study, the goal is to investigate educational videos returned by YouTube in terms of possible gender bias via objective measures. The evaluation is based on *group fairness* since it is investigated if the online materials are affected by societal stereotypes about gender in the context of education. Moreover in *group fairness*, *statistical parity*, *demographic parity* or more generally known as *equality of outcome* is specifically focused on, i.e. given a population divided into groups, the groups in the output of the system should be equally represented. In the scope of this chapter, *equality of outcome* is a more appropriate standard since *equal* gender representations is required in results. The main aim in this chapter is to detect bias with respect to *equality of outcome* using the *perceived* gender of narrators in videos returned by YouTube in response to educational queries comprising of keywords regarding some educational field. For this purpose, educational queries that are derived from the course modules of STEM and NON-STEM fields are used.

The main contributions can be summarised as follows:

- *Two new measures of bias* which are explained in Section 6.3.1 in detail are proposed that treat the two gender groups equal, and generate bias values which are symmetric and easy to interpret.
- The bias measures are implemented to *investigate possible perceived gender bias* for educational searches in YouTube about different majors from STEM and NON-STEM fields.
- Also *the relative bias is comparatively evaluated* for educational queries in YouTube from various majors from STEM and NON-STEM fields.
- Then, rank information is further incorporated into the bias analysis to investigate if various rank values affect first the *existence of bias*, then *difference in magnitude of bias* between STEM and NON-STEM fields as well as in the same field.

6.2 Preliminaries and Research Questions

Consider the following scenario: a query such as "Gravity" or "Python Programming" is submitted to YouTube, and a result page containing a collection of videos is returned in response. Such queries are called as educational queries and the abbreviation of *YVRP* is used for the YouTube video result page for a given query throughout this chapter. The gender of the narrator is explored in this chapter, and the first goal is to label the videos according to the narrator's *perceived* gender. The following values can be associated with a *perceived* gender label as *male*, *neutral*, *female*, *not-relevant*, and *N/A* with respect to the viewer's overall perception and their meanings are as follows:

- **male** (G_m) If the video is mostly narrated by people whose gender is perceived as male;
- **neutral** (G_{neut}) When the video does not favour either male or female gender in narration. Therefore, the video does not help the viewer to infer any gender dominance;
- **female** (G_f) If the video is mostly narrated by people whose gender is perceived as female;
- **not-relevant** (G_{not_rel}) when the video is not-relevant with respect to the educational query;
- **N/A** ($G_{N/A}$) When the annotation is not applicable for the video – the video is not in English or it has been removed from the system, or there is no narrator.

A *YVRP* contains 12 video links. In Figure 6.1, different ranked lists of labelled results are displayed. In Figure 6.1 (a) the perceived gender of all the narrators is labelled as male which demonstrates a clear bias. In Figure 6.1 (b) and (c) half of the perceived genders are male and half of them are female however in Figure 6.1 (b) the top 6 ranked videos are labelled as male while in Figure 6.1 (c) the top 6 are labelled as female. In Figure 6.1 (d) there is no obvious bias, while the videos of Figure 6.1 (e) has the labels of neutral, not-relevant, and N/A which further complicate the bias evaluation. Many and different queries must be issued first, and the results analysed. The first research question is:

RQ1: On a *perceived* male-female binary gender space, does YouTube return *biased* *YVRP*s in response to various educational queries?

TheThere are different fields of education which are broadly categorized as STEM and NON-STEM where the number of female students in some STEM fields has been considerably less than the males. Our second research question is:

RQ2: Is there a *significant* difference in *perceived* gender bias in *YVRP*s returned in response to STEM vs. NON-STEM educational queries?

We provide bias evaluation measures that take into account the rank of the results. One of the measure looks at the top n results in comparison to the rest of the 12 results, where n is the cut-off value. For example in the videos of Figure 6.1 (d)

Table 6.1 Symbols, functions, and labels used throughout this chapter

Symbols	
$\mathcal{Q}$	set of queries.
q	a query $q \in \mathcal{Q}$.
r	a ranked list of the given *YVRP* (list of retrieved videos).
r_i	the video in r retrieved at rank i.
$\|r\|$	size of r (number of videos in the ranked list).
n	number of videos considered in r (cut-off).
n_{mf}	number of videos in r which are annotated as male or female (excluding neutral, not-relevant, and N/A).
Functions	
$j(r_i)$	returns the label associated to r_i.
$f(r)$	an evaluation measure for *YVRP*s.
Labels	
G_m	*perceived* male gender.
G_{neut}	*perceived* neutral gender.
G_f	*perceived* female gender.
G_{not_rel}	not-relevant wrt a query.
$G_{N/A}$	N/A - gender annotation is not applicable.

where there is a cut-off value of 3, there will be a significant difference in bias in top 3 vs top 12. The third research question is:

RQ3: Do different cut-off values affect the existence of *perceived* bias and magnitude of bias difference between STEM and NON-STEM fields?

Finally, in addition to the impact of different cut-off values on the existence of *perceived* bias in STEM and NON-STEM fields, how the cut-off values influence on the magnitude of bias in each field is further examined. The last research question is:

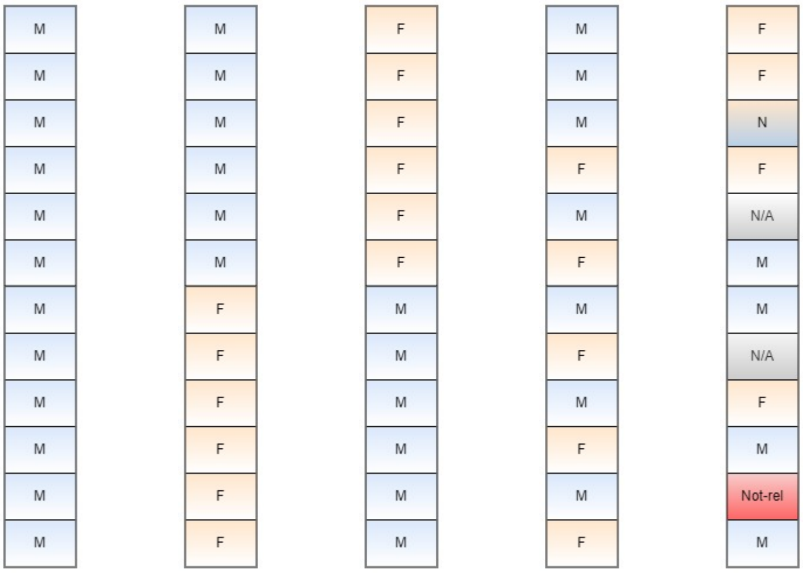
RQ4: Do different cut-off values affect the magnitude of *perceived* bias of STEM and NON-STEM fields?

6.3 Gender Bias Evaluation Methodology

This section describes the methodology for evaluating *perceived* gender bias using a binary gender assumption. Two bias measures are proposed and further a procedure is presented for identifying potential bias associated with those measures.

6.3.1 Measures of Bias

Figure 6.1 Example ranked lists of YouTube results for a query.



(a) List1	**(b)** List2	**(c)** List3	**(d)** List4	**(e)** List5
M	M	F	M	F
M	M	F	M	F
M	M	F	M	N
M	M	F	F	F
M	M	F	M	N/A
M	M	F	F	M
M	F	M	M	M
M	F	M	F	N/A
M	F	M	M	F
M	F	M	F	M
M	F	M	M	Not-rel
M	F	M	F	M

Let $\mathcal{Q}$ be the set of educational queries about major areas of study in STEM and NON-STEM fields. When a query $q \in \mathcal{Q}$ is issued to YouTube, YouTube returns a *YVRP* r. The *perceived* gender of the i-th retrieved video r_i with respect to q is defined as $j(r_i)$. For reference, Table 3.1 shows a summary of all the symbols, functions and labels used throughout this chapter.

For satisfying the group fairness criterion of *equality of outcome*, male and female genders should be *equally* represented in the retrieved YouTube videos. In the scope of *perceived* gender bias analysis, bias exists in a ranked list of videos retrieved by YouTube if the *perceived* gender representation *significantly deviates* from *equal* representation. Thus, the difference between the representation of two genders, namely as male and female, need to be measured.

Formally, the *perceived gender* bias in a *YVRP* r is measured as follows:

$$\Delta_f(r) = f_{G_m}(r) - f_{G_f}(r) \tag{6.1}$$

where f is a function that measures the likelihood of r in satisfying the information need of the user about the *perceived* gender of male (G_m) and female (G_f). When $\Delta_f(r) = 0$ we consider that r to be bias-free. When $\Delta_f(r) > 0$, the YVRP is biased towards male (G_m), with maximal bias when $\Delta_f(r) = 1$. When $\Delta_f(r) < 0$, then the YVRP is biased towards female (G_f), with maximal bias when $\Delta_f(r)$ = -1.

For the function $f(r)$, two *novel* bias measures are proposed in the scope of *equality* *of outcome*. Please note that only the videos annotated with the *perceived* gender labels of male (G_m) and female (G_f), that are relevant to the query, are taken into account. The videos for which $j(r_i)$ returns neutral (G_{neut}), not-relevant ($G_{not-rel}$), or N/A ($G_{N/A}$) are discarded. Note that, $j(r_i)$ returns the label of video r_i specifying its gender group. Based on this, $[j(r_i) = G_m]$ refers to a conditional statement which returns 1 if the video r_i is annotated as the member of G_m and 0 otherwise. These two new measures of *representation* and *exposure* are denoted by $\mathcal{R}ep@n$ and $\mathcal{E}xp@n$ respectively. The measure of $\mathcal{R}ep@n$ deals with the bias in *gender proportion*, while $\mathcal{E}xp@n$ aims to reveal the bias caused by *exposure effects*, i.e. attention received by ranked items. The first measure of bias, $\mathcal{R}ep@n$ which is interpreted with respect to the *perceived* gender label of *male* as follows:

$$\mathcal{R}ep_{G_m}@n = \frac{1}{n_{mf}} \sum_{i=1}^{n} [j(r_i) = G_m] \tag{6.2}$$

Note that $\mathcal{R}ep_{G_f}@n$ is computed in the same way. The following equation is obtained by substituting Eq. (6.2) in Eq. (6.1):

$$\Delta_{\mathcal{R}ep@n}(r) = \frac{1}{n_{mf}} \sum_{i=1}^{n} \Big([j(r_i) = G_m] - [j(r_i) = G_f] \Big)$$

Although the first bias measure of $\mathcal{R}ep@n$ is very intuitive, it is insensitive to the rank positions since all the search results in the first n documents contribute to the bias score equally, regardless of their rank positions. Thus, the second measure of $\mathcal{E}xp@n$ is proposed to address this issue by defining a discount function based on rank which includes a strong concept of ranking information in the bias analysis. The logarithmic discounting method is inspired by the weighted discount mechanism of nDCG (Järvelin & Kekäläinen, 2000) which is a widely used utility-based information retrieval metric. The proposed measure of $\mathcal{E}xp@n$ which is interpreted with respect to the *perceived* gender label of *male* as follows:

$$\mathcal{E}xp_{G_m}@n = \sum_{i=1}^{n} \frac{1}{\log(i+1)} \Big(\frac{[j(r_i) = G_m]}{[j(r_i) = G_m] + [j(r_i) = G_f]} \Big) \tag{6.3}$$

Note that $\mathcal{E}xp_{G_f}@n$ is computed in the same way. The following equation is obtained by substituting Eq. (6.3) in Eq. (6.1):

$$\Delta_{\mathcal{E}xp@n}(r) = \sum_{i=1}^{n} \frac{1}{\log(i+1)}\left(\frac{[j(r_i) = G_m] - [j(r_i) = G_f]}{[j(r_i) = G_m] + [j(r_i) = G_f]}\right)$$

The scores of the proposed measures are easy to interpret, for a given ranked list the scores of two gender groups sum up to 1. If the bias scores are interpreted with respect to the *equal* representation using $\mathcal{R}ep@n$, then it can be inferred which gender group is more/less represented than the desired representation. Same holds true for the exposure measure, $\mathcal{E}xp@n$, which determines if a gender group is more or less exposed than the desirable situation of the *equal* exposure. For interpreting the results, if the value of 0.5 which is the desired case, is subtracted from the measure scores of male and female gender for a given list, then the remaining bias scores of male and female are symmetric. Same holds for the exposure measure. These bias measures are calculated for the sample ranked lists in Figure 6.1 as follows:

- In Figure 6.1 (a), the perceived gender of all the narrators are labelled as male. For this ranked list r, $\mathcal{R}ep_{G_m}@12 = 1$, whereas $\mathcal{R}ep_{G_f}@12 = 0$, thereby $\Delta_{\mathcal{R}ep@12}(r) = 1$ which is the maximal bias. Same exposure bias score is computed for this ranked list.

- In Figure 6.1 (b), half of the perceived genders are male and the top 6 ranked videos are labelled as male. For this ranked list r, $\mathcal{R}ep_{G_m}@12 = 0.5$ and $\mathcal{R}ep_{G_f}@12 = 0.5$, thus $\Delta_{\mathcal{R}ep@12}(r) = 0$ indicating no representation bias. On the other hand, $\mathcal{E}xp_{G_m}@12 = 0.65$ and $\mathcal{E}xp_{G_f}@12 = 0.35$, thus $\Delta_{\mathcal{E}xp@12}(r) = 0.30$. Since the first measure only looks at the proportion of gender groups in the given ranked list without taking into account the rank information, no representation bias is observed. However, using the second measure which uses rank information with a logarithmic discount function, it can be seen that there exists an *exposure bias* towards the *male* gender since $\Delta_{\mathcal{E}xp@12}(r) > 0$.

- In Figure 6.1 (c), again half of the perceived genders are male however unlike Figure 6.1 (b), the top-6 ranked videos are labelled as female. For this ranked list r, our representation bias measure computes the same scores with the ranked list r in (b) as $\mathcal{R}ep_{G_m}@12 = 0.5$ and $\mathcal{R}ep_{G_f}@12 = 0.5$, thus $\Delta_{\mathcal{R}ep@12}(r)$ $= 0$ showing no representation bias. On the other hand, $\mathcal{E}xp_{G_m}@12 = 0.35$, whereas $\mathcal{E}xp_{G_f}@12 = 0.65$, thus $\Delta_{\mathcal{E}xp@12}(r) = -0.30$. As with the ranked list in (b), there does not exist representation bias since $\Delta_{\mathcal{R}ep@12}(r) = 0$, while there exists an *exposure bias* towards the *female* gender since $\Delta_{\mathcal{E}xp@12}(r) < 0$.

- In Figure 6.1 (d), the ranked list r contains almost the same number of male and female perceived gender labels, yet in the top-3 all the narrators are labelled as male. For this ranked list r, $\mathcal{R}ep_{G_m}@12 = 0.58$, whereas $\mathcal{R}ep_{G_f}@12 = 0.42$, thereby $\Delta_{\mathcal{R}ep@12}(r) = 0.16$ indicating a low representation bias, very close to the bias-free case. On the other hand, $\mathcal{E}xp_{G_m}@12 = 0.67$, whereas $\mathcal{E}xp_{G_f}@12 = 0.33$, thus $\Delta_{\mathcal{E}xp@12}(r) = 0.34$ revealing a higher bias in exposure than the representation. In this case, a higher exposure bias is observed bias since the exposure measure takes into account the rank information. Moreover, if bias is measured for different cut-off values, then various level of bias might be obtained in representation and exposure. For instance, let's only look at the top-3 positions in the ranked list r, where $n = 3$. Then, $\mathcal{R}ep_{G_m}@3 = 1$ and $\mathcal{R}ep_{G_f}@3 = 0$, thus $\Delta_{\mathcal{R}ep@3}(r) = 1$ which is the maximal bias in comparison to the full list. Similarly, $\Delta_{\mathcal{E}xp@3}(r) = 1.0$ which is the maximal bias for exposure as well.
- In Figure 6.1 (e), the ranked list r contains the same number of male and female perceived gender labels, yet female gender appears more in the top positions. Additionally, unlike the previous lists this ranked list contains neutral, not-relevant, and N/A. Since the perceived gender labels of neutral, not-relevant, and N/A do not contribute to detect gender bias, these labels are not included in our computations. Therefore, for this ranked list r, $\mathcal{R}ep_{G_m}@12 = 0.5$ and $\mathcal{R}ep_{G_f}@12 = 0.5$, thus $\Delta_{\mathcal{R}ep@12}(r) = 0$, no representation bias. This is because if the aforementioned labels are discarded, the ranked list turns into the same lists as in (b) and (c). On the other hand, $\mathcal{E}xp_{G_m}@12 = 0.35$, while $\mathcal{E}xp_{G_f}@12 = 0.65$, thus $\Delta_{\mathcal{E}xp@12}(r) = -0.30$ which indicates that there is an exposure bias. Similar to the ranked lists in (b) and (c), no representation bias is observed since $\Delta_{\mathcal{R}ep@12}(r) = 0$, whereas there exists an *exposure bias* towards the *female* gender since $\Delta_{\mathcal{E}xp@12}(r) < 0$.

These bias computations of the five different ranked lists demonstrate that both of the proposed measures are necessary since they provide different types of information for the *perceived* gender bias analysis. Moreover, the findings show that the magnitude of bias differs with different cut-off values (n), therefore bias is quantified using various values of n in Section 6.4.2. Since users tend to pay more attention to the top positions in search results, the impact of higher bias in these positions could be more severe in the scope of gender equality. In addition to the bias computations, for interpreting the bias scores in representation, the value of 0.5 can be subtracted to obtain the relative representations of male and female in a ranked list

r. For instance, in Figure 6.1 (d), $\mathcal{R}ep_{G_m}@12 = 0.58$ and $\mathcal{R}ep_{G_f}@12 = 0.42$ for male and female gender labels respectively. Thus, if the value of 0.5 is subtracted from those scores, the values of 0.08 for male and -0.08 for female are obtained, meaning that the male gender is represented 8% more, and the female gender is represented 8% less than the *equal* representation. Similarly, $\mathcal{E}xp_{G_m}@12 = 0.67$ and $\mathcal{E}xp_{G_f}@12 = 0.33$ for male and female gender labels respectively and if the value of 0.5 is subtracted, then the following values are obtained, 0.17 for the male and -0.17 for female genders. From this, it can be inferred that the male gender receives 17% more exposure while the female gender receives 17% less exposure than the desired case.

After the computation of representation and exposure bias scores, the mean bias (MB) and mean absolute bias (MAB) of these measures can be further computed over a set of queries in the dataset to aggregate the bias results. MB score of STEM field computes a mean value over all the STEM queries' scores for the corresponding measure, whereas MAB computes a mean value over all the absolute value of the measure scores for the STEM queries. Note that MB shows towards which *perceived* gender the results are biased and MAB solves the limitation of MB if different queries have bias contributions with opposite signs and cancel each other out. Thus, MB and MAB measures are complementary for aggregating the results and interpreting those results in a proper way.

Please note that in the scope of this chapter, the gender label of a given video is merely assigned based on the narrators' *perceived* gender and gender binary assumption is applied. However, the definitions and thereby the proposed measures of bias can easily be applied to studies where the gender label is defined in a more refined manner. Labels can also be assigned based on the male/female dominance, similar to the viewpoints presented by Draws, Tintarev, Gadiraju, Bozzon & Timmermans (2021). Yet, in the scope this gender bias study, the *perceived* gender label is adopted as binary for the preliminary results. Moreover, the proposed measures are also suitable for studies that use similar categorical features like age, education, ethnicity, or geographic location (Lipani et al., 2021) and seek for *demographic parity* specifically, in search settings.

6.3.2 Quantifying Bias

Using the measures of bias defined in Section 6.3.1, the *perceived* gender bias of the STEM and NON-STEM fields is quantified in YouTube videos returned in response to the educational queries in various majors, and further compared.

- **Collecting *YVRP*s.** The educational queries issued for searching in YouTube were obtained from *TheUniGuide* [1]. TheUniGuide is a free university advice service which is part of The Student Room [2] that helps students make more informed decisions about their higher education choices. Each query was submitted to YouTube using a UK proxy in *incognito mode* and crawled the top-12 video results returned by YouTube. Note that the data collection process was done in a controlled environment such that the queries were sent to YouTube by avoiding long time-lags. After having crawled all the video results related to the majors in both STEM and NON-STEM fields, they were labelled. The *perceived* gender label of each video was annotated with respect to the educational queries by analysing the gender(s) of the narrator(s) from the viewer's perspective.

- **Bias Evaluation.** The bias scores are computed for every *YVRP* with two novel bias measures with three different cut-off values: $\mathcal{R}ep@n$ and $\mathcal{E}xp@n$ for $n = 3, 6, 12$. Then, the results are aggregated using the MB and MAB. Additionally, first the existence of bias for each field is examined, further the bias results of STEM and NON-STEM fields with different measures and cut-off values are compared. Finally, the impact of different cut-off values is investigated on bias scores of STEM and NON-STEM fields.

- **Statistical Analysis.** To identify whether the bias measured is not due to noise, a one-sample t-test is applied: the null hypothesis is that no difference exists and that the true mean is equal to zero. Note that since the sample size is sufficiently large (> 30), according to the central limit theorem the sampling distribution is considered normal (Kwak & Kim, 2017). If this hypothesis is rejected, hence there is a significant difference and it can be claimed that the *YVRP*s of the evaluated field, STEM or NON-STEM is biased. The difference in bias measured across the two fields is further compared using a two-tailed *independent* t-test: the null hypothesis is that the difference between the two true means is equal to zero. If this hypothesis is rejected, hence there is a significant difference, then it can be claimed that there is a difference in bias between the two fields. The acceptance or rejection of the null hypothesis is fulfilled based on the p-values. Note that before applying the two-tailed independent t-test, the equality of variances such that if the two samples have equal or unequal variances is not checked, but rather independent t-

[1] https://www.theuniguide.co.uk/

[2] Free student discussion forum in UK

test with unequal variances is employed directly (Delacre, Lakens & Leys, 2017). Nonetheless, in the context of this gender bias analysis, it seems that independent t-test with equal or unequal variances do not make a noticeable difference in p-values based on the initial analysis. In addition to the statistical significance, namely p-values, effect sizes are also reported using Cohen's d. Statistical significance in the presented analysis helps to examine whether the findings show systematic bias or they are the result of noise, whereas effect sizes provide information about the magnitude of the differences which makes both p-values and effect sizes complementary for the interpretation of the results.

Apart from these, to investigate the effect of different cut-off values on bias results in the same field, STEM or NON-STEM, a two-tailed *paired* t-test is computed since in this analysis the same query set is examined only with different cut-off values. Moreover, Bonferroni correction is further applied (Sedgwick, 2012) for multiple hypothesis testing since there are 24 hypotheses in total in the context of cut-off value analysis. Thus, without the Bonferroni correction, with the significance level, $\alpha = .05$ and 24 hypotheses, the probability of identifying at least one significant result due to chance is around 0.71 which means that the results could be misleading. Hence, the Bonferroni correction is also applied for more reliable results in the scope of the cut-off value analysis in Section 6.4.2. Note that for the significance level where $\alpha = .05$, and with the Bonferroni correction new $\alpha = .002$. Thus, Bonferroni correction rejects the null hypothesis for each p-value (p_i) if $p_i <= .002$ instead of .05.

6.4 Experimental Setup

This section provides the description of the experimental setup based on the proposed method as defined in Section 6.3. Initially the information about the dataset is provided, further the details about the annotation process is given. Lastly, the *per-*

[3] Total War in the modern era

[4] Fundamentals of Design

[5] PC technology

[6] Human behaviour

[7] Transition to work

Table 6.2 All the course modules of TheUniGuide we used as the user queries for the main study.

STEM	Course Modules			NON-STEM	Course Modules		
Biology	Biochemistry	Evolution and biodiversity	Marine and terrestrial ecology	**English Language and Literature**	Explorations in literature	Chaucer: texts, contexts, conflicts	Shakespeare in performance english language and literature
	Plant science in biology	Human physiology	Habitat ecology in biology		Renaissance literature	Modernist fiction	Creative writing: drama
	Environmental issues	Molecular methodology for biologists	Cell structure and function		British romanticism	Literary and cultural theory	Stylistics in literature
	Principles of genetics				Aspects of modernism in literature		
Chemistry	Solid state chemistry	Shapes, properties and reactions of molecules	Organic and biological chemistry	**Politics**	Central themes in political thought	Modern British politics	Capital labour and power: Britain 1707-1939
	Chemistry for the physical sciences	Molecular pharmacology	States of matter in chemistry		The holocaust in politics	War in the industrial age politics [3]	Freedom, power and resistance: an introduction to political ideas
	Chemistry of materials	Inorganic chemistry	The global Earth system		International politics	Making of the modern world in politics	The political economy of development
	Mineralogy and petrology (typo exists in the original query)				Comparing extremism in European liberal democracies		
Computer Science	Organisational behaviour in practice	Principles of programming	Data management in computer science	**Psychology**	Cell biology in psychology	Mind and behaviour	Exploring effective learning in psychology
	Mathematics for computer science	Languages and computability	Fundamentals of Computer Design [4]		Experimental methods and statistical	Individual and social processes	Development psychology
	Personal Computer technology [5]	Image processing	Software systems development in computer science		Brain and cognition in psychology	Social psychology	Humans in biological perspective in psychology
	Human computer interaction				Evolution and behaviour in psychology		
Mathematics	Calculus	Algebra	Structured programming	**Public Relations**	Business strategy	Internal corporate communication	Social media or public relations
	Algorithms and applications	Coordinate and vector geometry	Differential equations		Work and organisational change	Behavioural science [6]	Management in context
	Probability	Regression and anova	Analytical and computational foundations in maths		Work experience in public relations [7]	Business fundamentals	Managing the brand
	Problem solving methods in maths				Design in marketing		
Physics	Laboratory physics	Contemporary physics	Mathematical techniques in physics	**Sociology**	Observing in sociology	Urban sociology	Understanding deviance and social problems in sociology
	Quantum physics	Newtonian and relativistic mechanics	Fabric of physics		Individual and society	Applied ethics	Media and crime in sociology
	Plasma and fluids in physics	Special and general relativity	Analysing the nanoscale and magnetism		Nature and society in sociology	Sexuality and social control in sociology	Contemporary work and organisational life in sociology
	Stellar physics				Mobilisation, social movements and protest in sociology		

ceived gender bias results are displayed and discussed.

6.4.1 Dataset

Throughout this chapter, the main aim is to mimic a user scenario in which the user is trying to decide on his/her major through searching educational queries or course modules of various majors on YouTube. Thus, the study is designed accordingly and all the educational queries were obtained from *TheUniGuide*. The first reason for choosing TheUniGuide is that when a query of "university chemistry courses" is searched in *incognito mode* with a UK proxy to construct the set of educational queries for *chemistry*, TheUniGuide appears as the top result in Google search. Second, the other search results were mainly the official pages of different universities about the corresponding major and the curriculum information of specific universities deliberately were not selected to create the dataset of courses which are used as queries. Third, the web pages for different majors from STEM and NON-STEM fields were examined and it was observed that TheUniGuide website provides comprehensive information about a major [8].

5 STEM majors of chemistry, physics, biology, maths, and computer science and 5 NON-STEM majors of sociology, psychology, politics, public relations, and English language and literature were selected. Those majors were chosen since it is believed that they sufficiently span distinct areas in STEM and NON-STEM fields which might have different male/female gender proportions. For the crawling, a specific scenario was simulated such that the user is a prospective university student who uses YouTube for deciding on his/her major in STEM and NON-STEM fields. Initially, the course modules of each selected major were crawled from TheUniGuide. The UK proxy and the YouTube desktop version in *incognito mode* were used then the region was set as the UK and language as English automatically, other settings were left as default. Note that in the filter options at the top, "sort by" option was selected as "relevance" by default which means that the search results will be ranked based on relevance.

In the scope of this chapter, since personalised search might complicate the bias analysis, the analysis was designed in unpersonalised search settings. The study was initially designed as follows. In the crawling process, top-12 relevant videos and recommended videos by Youtube were included. The videos of the top-3 query recommendations by YouTube were crawled as well. However, the recommended video results and *YVRP*s in response to the query recommendations did not contain context-specific results about the issued query. Then, the dataset crawling process was modified since those video results would probably not attract the user's attention, i.e. the user is a prospective university student searching for information about different majors, in a real world scenario. For these reasons, in the modified dataset crawling process, only relevant *YVRP*s for a given query were crawled and all the course modules were obtained from TheUniGuide and used as user queries. Note that TheUniGuide has 10 course modules for each major which can be used as queries, hence both STEM and NON-STEM fields have 50 queries and ranked lists in total. In the modified data crawling process, only the 12 relevant video results returned by YouTube per query were crawled. In this way, each course module/query contained 12 video URLs and each major had 120, thereby the dataset consisted of 1200 video URLs in total.

In addition to these, for some queries YouTube did not return relevant video results in the educational context, therefore the queries/course modules were slightly modified and these modifications are indicated by red color in Table 6.2. Note that

[8] For the major of "chemistry", please go to `https://www.theuniguide.co.uk/subjects/chemistry`.

Table 6.3 Retrieval performance of YouTube, p-values of a two-tailed independent t-test computed between STEM and NON-STEM fields

	P@12	DCG@12	RBP
STEM	0.9450	4.8306	0.8896
NON-STEM	0.9483	4.8475	0.8870
p-value	.87	.87	.89

Table 6.4 *Perceived* gender bias in YouTube for the top-12 relevant results, p-values of a two-tailed independent t-test computed between STEM and NON-STEM fields

		P@12	DCG@12	RBP
MB	STEM	**0.5200*****	**2.7032*****	**0.5065*****
	NON-STEM	**0.3117*****	**1.5636*****	**0.2828*****
	p-value	**.0064**	**.0035**	**.0023**
	effect size d	**0.564**	**0.606**	**0.633**
MAB	STEM	0.5400***	**2.8125*****	**0.5274*****
	NON-STEM	0.4450***	**2.1368*****	**0.3991*****
	p-value	.0897	**.0255**	**.0221**
	effect size d	0.346	**0.458**	**0.470**

the majority of the queries were adopted as they are written in TheUniGuide with lowercase/uppercase letters as well as punctuation symbols to avoid injecting any personal bias. For the rest of the queries, if YouTube did not return context-specific video results for the original query, i.e. if the query is related to *sociology* then in the sociology-context, then the original query were slightly changed as follows. The query was expanded only by adding the context/major field information as displayed in Table 6.2. However, for only 5 queries in total, this solution was not sufficient so for these queries the query itself was changed to specify the context properly. For the queries of *computer science*, the abbreviation of *PC* in the first query were written in its full form, and only the term *Computer* was added to the second query. For the NON-STEM queries, specifying the context was even more difficult since those queries were related to a wide-range of topics. For instance, for the original query of *Human behaviour*, YouTube returned the music videos of Björk, an Icelandic singer. Therefore, three queries in total were paraphrased to obtain the relevant instructional video results as highlighted in Table 6.2. Note that before the paraphrasing, sufficient information related to those queries was collected from the Web to change the original queries properly for the context-specific results. By designing the dataset crawling process this way, the search scenario turns out to be more realistic in an educational context which can better detect *perceived* gender bias that the user is exposed to in real world. For the sake of reproducibility, the annotated dataset is publicly available at `https://github.com/gizem-gg/Youtube-Gender-Bias`.

As for the annotation procedure, a video in a *YVRP* is annotated based on its

Table 6.5 *Perceived* gender bias in YouTube for the top-12 relevant results, p-values of a two-tailed independent t-test computed between STEM and NON-STEM fields

		$\mathcal{R}ep$@3	$\mathcal{R}ep$@6	$\mathcal{R}ep$@12	$\mathcal{E}xp$@3	$\mathcal{E}xp$@6	$\mathcal{E}xp$@12
MB	STEM	**0.6200*****	**0.5460*****	**0.5558*****	**0.6012*****	**0.5541*****	**0.5600*****
	NON-STEM	**0.3067*****	**0.2820*****	**0.3266*****	**0.3083*****	**0.2880*****	**0.3144*****
	p-value	**.0044**	**.0050**	**.0049**	**.0108**	**.0047**	**.0023**
	effect size d	**0.590**	**0.581**	**0.583**	**0.525**	**0.584**	**0.632**
MAB	STEM	**0.7000*****	0.5940***	0.5794***	**0.7136*****	**0.6225*****	**0.5877*****
	NON-STEM	**0.5467*****	0.4767***	0.4808***	**0.5566*****	**0.4683*****	**0.4449*****
	p-value	**.0355**	.0768	.0855	**.0334**	**.0171**	**.0174**
	effect size d	**0.431**	0.361	0.351	**0.436**	**0.490**	**0.489**

Table 6.6 *Perceived* gender bias in YouTube for the top-12 relevant results, p-values of a two-tailed paired t-test computed between STEM and NON-STEM fields

		STEM	NON-STEM		STEM	NON-STEM		STEM	NON-STEM
MB	$\mathcal{R}ep$@3	0.6200***	0.3067***	$\mathcal{R}ep$@6	0.5460***	0.2820***	$\mathcal{R}ep$@3	0.6200***	0.3067***
	$\mathcal{R}ep$@6	0.5460***	0.2820***	$\mathcal{R}ep$@12	0.5558***	0.3266***	$\mathcal{R}ep$@12	0.5558***	0.3266***
	p-value	.11	.62		.79	.20		.26	.77
	effect size d	0.082	0.046		-0.026	-0.096		0.490	-0.039
	$\mathcal{E}xp$@3	0.6012***	0.3083***	$\mathcal{E}xp$@6	0.5541***	0.2880***	$\mathcal{E}xp$@3	0.6012***	0.3083***
	$\mathcal{E}xp$@6	0.5541***	0.2880***	$\mathcal{E}xp$@12	0.5600***	0.3144***	$\mathcal{E}xp$@12	0.5600***	0.3144***
	p-value	.19	.59		.83	.31		.39	.91
	effect size d	0.098	0.038		-0.015	-0.058		0.093	-0.012

Table 6.7 *Perceived* gender bias in YouTube for the top-12 relevant results, p-values of a two-tailed paired t-test computed between STEM and NON-STEM fields

		STEM	NON-STEM		STEM	NON-STEM		STEM	NON-STEM
MAB	$\mathcal{R}ep$@3	0.7000***	0.5467***	$\mathcal{R}ep$@6	0.5940***	0.4767***	$\mathcal{R}ep$@3	0.7000***	0.5467***
	$\mathcal{R}ep$@6	0.5940***	0.4767***	$\mathcal{R}ep$@12	0.5794***	0.4808***	$\mathcal{R}ep$@12	0.5794***	0.4808***
	p-value	.0156	.11		.66	.89		.0132	.22
	effect size d	0.299	0.215		0.045	-0.014		0.369	0.210
	$\mathcal{E}xp$@3	0.7136***	0.5566***	$\mathcal{E}xp$@6	0.6225***	0.4683***	$\mathcal{E}xp$@3	**0.7136*****	0.5566***
	$\mathcal{E}xp$@6	0.6225***	0.4683***	$\mathcal{E}xp$@12	0.5877***	0.4449***	$\mathcal{E}xp$@12	**0.5877*****	0.4449***
	p-value	.0071	.0132		.15	.36		**.0014**	.0209
	effect size d	0.269	0.261		0.114	0.077		**0.389**	0.336

relevancy with respect to the given query as *relevant*, *not-relevant*, or *N/A*. If the video is *relevant* to the given query, then its narrators' gender perceived by the viewer is annotated using the gender labels of *male*, *female*, or *neutral.* Before the annotation of the actual dataset, two annotators initially annotated the dataset of top-12 relevant *YVRP*s that were crawled for the first user study. Then, the design of the gender bias study, thus the data crawling procedure were modified. Nonetheless,

Table 6.8 *Perceived* gender bias for specific majors of STEM and NON-STEM fields in YouTube for the top-12 relevant results - red denotes bias towards male while blue towards female

	Biology	*Chemistry*	*CS*	*Maths*	*Physics*	*Eng. Lan. Lit.*	*Politics*	*Psychology*	*Pub. Rel.*	*Sociology*
$\mathcal{R}ep$@3	**0.4333**	0.5667	**0.7667**	0.7333	0.5667	**0.6333**	0.3667	0.3000	**-0.1667**	0.4333
$\mathcal{R}ep$@6	**0.2833**	0.5033	0.5533	**0.7667**	0.6067	**0.7033**	0.3833	0.3400	**-0.1667**	0.1667
$\mathcal{R}ep$@12	**0.2061**	0.5699	0.5930	0.6833	**0.7098**	**0.7722**	0.3701	0.3619	**-0.0466**	0.1809
$\mathcal{E}xp$@3	**0.4524**	0.5235	0.7366	**0.7631**	0.5007	**0.5831**	0.3765	0.2656	**-0.1696**	0.5088
$\mathcal{E}xp$@6	**0.3523**	0.4846	0.5858	**0.7762**	0.5527	**0.6446**	0.3857	0.3022	**-0.1681**	0.2908
$\mathcal{E}xp$@12	**0.2736**	0.5451	0.6031	**0.7119**	0.6478	**0.7138**	0.3781	0.3241	**-0.0917**	0.2560

inter-rater agreement score which is calculated as a percentage of agreement between two annotators was computed. Pairwise agreement was examined; enter 1 if there is agreement and 0 if there is no agreement. After that, the mean of the fractions was calculated and the inter-rater agreement score for the initial study on the top-12 relevant *YVRP*s was over 0.90. Since the inter-rater agreement score is high, it shows that the annotation procedure does not prone to disagreements due to the simplicity of the task and does not require expert knowledge. Thus, the labelling has been fulfilled using a single annotator. Conditions of annotating a video with these labels are detailed in Section 6.2.

6.4.2 Results

Initially, the proposed bias measures for measuring the stance/ideological bias in Chapter 3 were computed on the new dataset to see whether they give consistent results with the new measures in the context of *perceived* gender bias. Prior to examining the possibility of bias in *YVRP*, first YouTube's retrieval performance was examined for queries/course modules from STEM and NON-STEM fields. In Table 6.3 it is observed that the retrieval performance of YouTube is high for the course modules coming from both STEM and NON-STEM fields. The retrieval performance for the queries/course modules in NON-STEM seems to be slightly better (for the first two measures); but their difference is statistically not significant. This is verified across all three IR evaluation measures. Following that, it is determined whether YouTube returns biased results in terms of the narrator's *perceived gender* for STEM and NON-STEM fields separately, and if so, whether YouTube's top-12 relevant search results suffer from the same level of bias, such that the difference

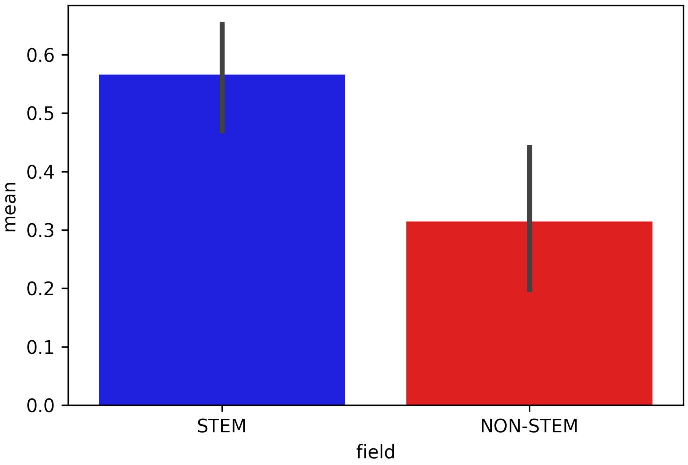


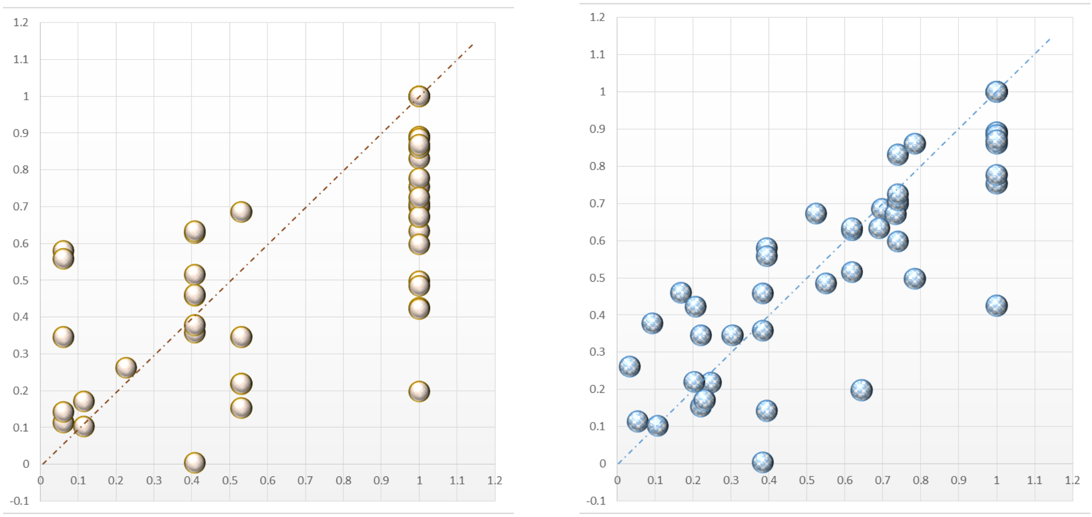
Figure 6.2 MB scores of $\Delta_{\mathcal{E}xp@12}$ measured on *perceived* gender labels of STEM and NON-STEM fields.

between STEM and NON-STEM fields is statistically not significant.

In Table 6.4, all MB scores are positive, STEM and NON-STEM fields seem to be biased towards the same *perceived* gender which is *male*. The one-sample t-test was applied on MB scores to check the existence of *perceived* gender bias, i.e. if the true mean is different from zero. The results show that these biases are statistically significant with p-value $< .0001$ denoted as *** in Table 6.4. Comparing STEM and NON-STEM fields on MB scores, it is observed that their differences are statistically significant and it is shown with the two-tailed *independent* t-test on P@12, DCG@12, and RBP. Note that the differences are statistically significant with difference confidence values, i.e. p-value $= .0064$, $.0035$, $.0023$ for P@12, DCG@12, and RBP respectively. In addition to the p-values, the corresponding effect sizes using Cohen's d were reported as well. Statistical significance, namely p-values in the analysis help to examine whether the findings show systematic bias or they are the result of noise, whereas effect sizes provide information about the magnitude of the differences which makes both p-values and effect size information complementary for the interpretation of the presented results. Based on MAB scores, it is observed that both STEM and NON-STEM suffer from an absolute bias. The difference between STEM and NON-STEM which is shown with the two-tailed *independent* t-test is statistically not significant for P@12 while statistically significant for DCG@12 and RBP with different confidence values, i.e. p-value $= .0255$, $.0221$ respectively.

In Table 6.5, the *perceived* gender bias is displayed using the measures proposed in Section 6.3.1, namely $\mathcal{R}ep@n$ and $\mathcal{E}xp@n$ for different cut-off values of $n = 3, 6, 12$. All MB and MAB scores are positive for both bias measures; the one-sample t-test computed on MBs and MABs are statistically significant for the measures where

Figure 6.3 MAB scores of $\Delta_{\mathcal{E}xp@n}$ measured on *perceived* gender labels of STEM field.



(a) $\Delta_{\mathcal{E}xp@n}$ measured on *perceived* gender labels, where x-axis is $n = 3$ and y-axis is $n = 12$.

(b) $\Delta_{\mathcal{E}xp@n}$ measured on *perceived* gender labels, where x-axis is $n = 6$ and y-axis is $n = 12$.

p-value < .0001 denoted as ***.

The two-tailed *independent* t-test computed on MBs to compare the difference in bias between the STEM and NON-STEM fields, the results indicate that their differences are statistically significant on the bias measures of $\mathcal{R}ep@n$ and $\mathcal{E}xp@n$ where n = 3, 6, 12, yet with different confidence values. The difference is statistically significant on $\mathcal{R}ep@3$, $\mathcal{R}ep@6$, and $\mathcal{R}ep@12$ with p-value = .0044, .0050, .0049, while p-value = .0108, .0047, .0023 on $\mathcal{E}xp@3$, $\mathcal{E}xp@6$, and $\mathcal{E}xp@12$ respectively. Some effect sizes that correspond to the difference of bias using MB scores are negative which indicates that the MB score of the *perceived* gender group of female is higher than male, albeit statistically not significant. For the MAB scores, the difference between STEM and NON-STEM fields is statistically significant on $\mathcal{R}ep@3$ with p = .0355 and statistically not significant where n = 6, 12. On the bias measure of $\mathcal{E}xp@n$, the differences based on MAB scores are statistically significant for all three cut-off values with p-value = .0334, .0171, .0174 where n = 3, 6, 12 respectively.

Regarding the impact of different cut-off values, in Table 6.6 using MB scores, it is observed that, both STEM and NON-STEM fields show similar scores for different cut-off values on both $\mathcal{R}ep@n$ and $\mathcal{E}xp@n$ measures, i.e. the two-tailed paired t-test computed on MB scores are statistically not significant. On the other hand, in Table 6.7 using MAB scores, it is observed that cut-off values might affect the *perceived* gender bias in STEM field. The two-tailed paired t-test computed on MABs of STEM field is statistically significant only for the measure of $\mathcal{E}xp@n$ between the following cut-off values, $n = 3$ vs. $n = 12$ with p-value = .0014 which corresponds to the significance level of $\alpha = .05$ after Bonferroni correction was applied.

In Table 6.8, the bias scores for each major in STEM and NON-STEM fields are displayed using the measures $\mathcal{R}ep@n$ and $\mathcal{E}xp@n$ for different cut-off values. Note that the highest/lowest bias scores are denoted as highlighted. Figure 6.2 displays the comparison of the bias scores in STEM and NON-STEM fields. The error bars show the standard error on the scores of the corresponding field. In Figure 6.3 (a) and (b), the overall *perceived* gender bias scores are compared on MAB scores ($\Delta_{\mathcal{E}xp@3}$ and $\Delta_{\mathcal{E}xp@12}$, $\Delta_{\mathcal{E}xp@6}$ and $\Delta_{\mathcal{E}xp@12}$ respectively), i.e. difference between the male and female gender scores, of *YVRP*s for each educational query in STEM field.

6.5 Concluding Discussion

Initially, it is verified if the *YVRP*s are biased using the proposed representation and exposure measures (**RQ1**). If so, then it is investigated if the *YVRP*s returned in response to the educational queries of STEM and NON-STEM fields suffer from the different level of bias (**RQ2**) by examining if the difference between the bias scores of the corresponding *YVRP*s are statistically significant. In Table 6.4, using the proposed measures in Chapter 3, all MB scores are positive and regarding the **RQ1**, the *YVRP*s of STEM and NON-STEM fields are both biased. The one-sample t-test was applied on MB scores to check the existence of bias, i.e. if the true mean is different from zero, if not this means that the difference appears due to noise, as explained in Section 6.3.2. These biases are statistically significant; there exists a systematic gender bias, i.e. preference of one gender over another, with p-value $<$ .0001. These results indicate that both STEM and NON-STEM fields are biased towards the *male* gender (all MB scores are positive). Based on MAB scores, it is observed that the *YVRP*s of both STEM and NON-STEM fields suffer from an absolute bias. Similar to Table 6.4, in Table 6.5, regarding the **RQ1** the *YVRP*s of STEM and NON-STEM fields are both biased. These findings suggest that both STEM and non-STEM fields are biased towards male (all MB scores are positive). On the basis of MAB scores, it is observed that both STEM and NON-STEM exhibit an absolute bias. With respect to the **RQ3**, it is examined whether different cut-off values affect the presence of bias, and the findings reveal that both STEM and NON-STEM fields are biased regardless of the cut-off values. Note that both groups of measures in Table 6.4 and Table 6.5 show consistent results; yet the new bias measures provide bias results that are easy to interpret which is more important in the context of *perceived* gender bias.

Regarding the **RQ2**, STEM and NON-STEM fields show different magnitude of bias – the two tailed *independent* t-tests applied on MB scores in Table 6.5 are statistically significant for $\mathcal{R}ep@n$ and $\mathcal{E}xp@n$ where $n = 3, 6, 12$. Yet, the two tailed *independent* t-tests applied on MAB scores are statistically significant for $\mathcal{R}ep@n$, only where $n = 3$ and $\mathcal{E}xp@n$ where $n = 3, 6, 12$. These findings indicate that generally STEM is more biased than NON-STEM field. Regarding the **RQ3**, cut-off values do not affect the existence of bias since the one-sample t-test computed on MBs and MABs are statistically significant where p-value $< .0001$ denoted as *** for the measures irrespective of the cut-off values.

Regarding the **RQ4**, it is investigated whether different cut-off values change the magnitude of bias – the two tailed *paired* t-tests applied on MB scores are statistically not significant regardless of the measure and cut-off value, see Table 6.6. Unlike the MB scores, the two tailed *paired* t-tests applied on MAB scores show statistically significant results only for the $\mathcal{E}xp_{prob}@3$ and $\mathcal{E}xp_{prob}@12$ of the STEM field. These findings suggest that generally cut-off values do not affect the magnitude of bias that STEM or NON-STEM field exhibit.

In Table 6.8 it is investigated which majors show the highest/lowest bias in STEM and NON-STEM fields. The empirical findings indicate that *Biology* is biased on both measures towards the male gender which shows the lowest bias score in STEM field and different cut-off values do not affect this. Unlike *Biology*, different measures and cut-off values change the major which shows the highest bias score towards the male gender. For the exposure measure, *Mathematics (a.k.a Maths)* is also biased towards the male gender and shows the highest bias score regardless of the cut-off values. On the other hand, for the representation measure the major with the highest bias score depends on the cut-off value. For $n = 3$, *Computer Science (a.k.a CS)* is the major showing the highest bias score, while for $n = 6$, *Mathematics (a.k.a Maths)* similar to the exposure measures, and for the full list *Physics* shows the highest score, where $n = 12$. Nonetheless for the NON-STEM field, the majors showing highest/lowest bias scores change neither with different measures nor cut-off values. Among the majors in NON-STEM field, *Public Relations (a.k.a Pub. Rel.)* shows the lowest bias which is the only major that is biased towards the female gender, whereas *English Language and Literature (a.k.a Eng. Lan. Lit.)* shows the highest bias score which is biased towards the male gender like majority of the majors in STEM and NON-STEM fields.

Moreover, the majors showing the highest bias in STEM field seem to be more biased in absolute value (magnitude) on average than their counterparts in NON-STEM field. Similarly, *Biology* shows higher bias scores on average than *Public Relations (a.k.a Pub. Rel.)* when their scores are compared in absolute terms. These results seem to be consistent with our aforementioned findings in Table 6.5 that STEM field is more biased which is towards the male gender. Also, these empirical findings in Table 6.8 support the implication that the *YVRP*s of some majors in NON-STEM field are biased towards the male, while others towards the female gender. Furthermore, the bias scores of majors in STEM field are more similar; standard deviation is smaller than the majors in NON-STEM field, i.e. average standard deviation is 0.15 for STEM majors while 0.27 for NON-STEM. Finally, looking at the major-specific bias scores we can observe that there exists a noticeable bias towards the male gender, even if *Public Relations (a.k.a Pub. Rel.)* seems to be biased towards the female gender, yet unlike STEM majors it does not show a strong bias.

In Figure 6.2, the results show that both STEM and NON-STEM are overall biased towards the male (positive mean scores) but STEM is more biased. The error bar of NON-STEM is slightly higher than STEM field. Figure 6.3 (a) and (b) displays the MAB scores of the measure,$\mathcal{E}xp@n$ for the STEM field. It is observed that there is a magnitude of bias difference between $n = 3$ and $n = 12$ (Figure 6.3a) while there is no visible difference between $n = 6$ and $n = 12$ (Figure 6.3b). These observations are consistent with the findings in Table 6.7 that only the difference between $\mathcal{E}xp@3$ and $\mathcal{E}xp@12$ is statistically significant for the STEM field.

7. INVESTIGATING THE SOURCE OF GENDER BIAS IN ONLINE EDUCATION

7.1 Introduction

Similar to the methodology in Section 6.2, in the context of this chapter same educational queries will be used for the bias analysis and the abbreviation of *YVRP* is used for the YouTube video result page. The gender of the narrator is similarly explored in this chapter, and the main aim is to label the videos according to the narrator's *perceived* gender. Unlike, a given video will not be labelled with a single gender label rather a probability distribution of male and female labels that could depict a more realistic *perceived* gender spectrum of a given video. Note that the *perceived* gender annotation will be automatically determined using the voice information of narrator(s) instead of manual annotation. For the detailed annotation procedure, please refer to Section 7.3.2.

Using an automated approach to obtain the *perceived* gender probability distributions of videos through voice information, this chapter mainly aims to following research questions. The first research question is:

RQ1: On a *perceived* male-female gender **spectrum**, does YouTube return *biased YVRP*s in response to various educational queries **using more fine-grained measures**?

The second research question is:

RQ2: Is there a *significant* difference in *perceived* gender bias that is computed in a **more fine-grained manner** in *YVRP*s returned in response to STEM vs. NON-STEM educational queries?

The third research question is:

RQ3: Do different cut-off values affect the existence of *perceived* bias and magnitude of bias difference by **using more fine-grained measures** between STEM and NON-STEM fields?

The fourth research question is:

RQ4: Do different cut-off values affect the magnitude of *perceived* bias of STEM and NON-STEM fields separately that is measured **in a more fine-grained manner**?

The last research question which is totally a new question that could not be answered in the scope of Chapter 6. Finally, in the context of this chapter, the source of bias (if exists) will be investigated. If the bias measured in the full *YVRPs* are consistent with the top video search results, i.e. especially in top-3, top-10 since these search results attract users' attention the most, then it can be inferred that the bias comes from the data itself. If there are some differences between those bias results, then the ranking algorithm could also be blamed. Note that the data and ranking algorithm both could be responsible for the biased *YVRPs*.

The fifth research question is:

RQ5: What is **the source of bias (if exists)**, does it come from the input data, or the ranking algorithm?

7.2 Gender Bias Evaluation Methodology

This section describes the methodology for evaluating *perceived* gender bias without a binary gender assumption. The bias measures proposed in 6.3.1 will be adapted to using male/female probability values instead of single *perceived* gender labels. Further, a similar evaluation procedure presented in 6.3.1 will be fulfilled for identifying potential bias as well as tracking the source of bias (if applicable) associated with those adapted measures.

7.2.1 Measures of Bias

Table 7.1 Symbols, functions, and labels used throughout this chapter

Symbols	
$\mathcal{Q}$	set of queries.
q	a query $q \in \mathcal{Q}$.
r	a ranked list of the given *YVRP* (list of retrieved videos).
r_i	the video in r retrieved at rank i.
$\lvert r \rvert$	size of r (number of videos in the ranked list).
n	number of videos considered in r (cut-off).
$prob_{mf}$	sum of probability scores of the videos in r associated with only male or female (not-relevant, N/A discarded).
Functions	
$prob_{G_m}(r_i)$	returns the probability value corresponds to male gender for r_i.
$prob_{G_f}(r_i)$	returns the probability value corresponds to female gender for r_i.
$f(r)$	an evaluation measure for *YVRP*s.
Labels	
G_m	*perceived* male gender.
G_f	*perceived* female gender.
G_{not_rel}	not-relevant wrt a query.
$G_{N/A}$	N/A - gender annotation is not applicable.

Let $\mathcal{Q}$ be the set of educational queries about the majors in STEM and NON-STEM fields. When a query $q \in \mathcal{Q}$ is issued to YouTube, YouTube returns a *YVRP* r. The probability value associated with the *perceived* gender of the i-th retrieved video r_i with respect to q is defined as $prob_{G_m}(r_i)$ for male and $prob_{G_f}(r_i)$ for female. For reference, Table 3.1 shows a summary of all the symbols, functions and labels used throughout this chapter.

In the scope of this chapter, similar to Section 6.3.1, the main aim is to satisfy group fairness criteria of *equality of outcome* where male and female genders should be equally represented in *YVRPs*. Thus the *perceived* gender bias is measured as the difference between the representation of male and female genders.

Formally, the *perceived gender* bias in a *YVRP* r is measured as follows:

$$\Delta_f(r) = f_{G_m}(r) - f_{G_f}(r) \tag{7.1}$$

For the function $f(r)$, two bias measures that are proposed in 6.3.1 are adapted to using probability scores instead of single *perceived* gender labels in the scope of this chapter. Note that the videos annotated with G_{not_rel} and $G_{N/A}$ are initially discarded before the bias score computations. The two adapted measures of *representation* and *exposure* are denoted by $\mathcal{R}ep_{prob}@n$ and $\mathcal{E}xp_{prob}@n$ respectively. The first adapted measure of bias, $\mathcal{R}ep_{prob}@n$ which computes a bias score using probability values associated with the *perceived* gender label of *male* as follows:

$$\mathcal{R}ep_{prob}G_m@n = \frac{1}{prob_{mf}}\sum_{i=1}^{n} prob_{G_m}(r_i) \tag{7.2}$$

Note that $\mathcal{R}ep_{prob}G_f@n$ is computed in the same way. The following equation by substituting Eq. (7.2) in Eq. (7.1):

$$\Delta_{\mathcal{R}ep_{prob}@n}(r) = \frac{1}{prob_{mf}}\sum_{i=1}^{n}\Big(prob_{G_m}(r_i) - prob_{G_f}(r_i)\Big)$$

Since the first bias measure of $\mathcal{R}ep_{prob}@n$ has a weak sense of rank information, i.e. all the positions contribute to bias score in an equal manner, the second measure is presented by adapting $\mathcal{E}xp@n$ that was proposed in Section 6.3.1 to using probability scores. The second adapted measure of bias, $\mathcal{E}xp_{prob}@n$ which computes a bias score in terms of exposure, using probability values associated with the *perceived* gender label of *male* as follows:

$$\mathcal{E}xp_{prob}G_m@n = \sum_{i=1}^{n}\frac{1}{\log(i+1)}(\frac{prob_{G_m}(r_i)}{prob_{G_m}(r_i)+prob_{G_f}(r_i)}) \tag{7.3}$$

Note that $\mathcal{E}xp_{prob}G_f@n$ is computed in the same way. The following equation is obtained by substituting Eq. (7.3) in Eq. (7.1):

$$\Delta_{\mathcal{E}xp_{prob}@n}(r) = \sum_{i=1}^{n}\frac{1}{\log(i+1)}(\frac{prob_{G_m}(r_i)-prob_{G_f}(r_i)}{prob_{G_m}(r_i)+prob_{G_f}(r_i)})$$

The scores of the proposed measures are easy to interpret, for a given ranked list, the scores of two gender groups sum up to 1. If the bias scores are interpreted with respect to the *equal* representation using $\mathcal{R}ep_{prob}@n$, then it can be inferred which gender group is more/less represented than the desired representation. Same holds true for the exposure measure, $\mathcal{E}xp_{prob}@n$, which determines if a gender group is more or less exposed than the desirable situation of the *equal* exposure. For interpreting the results, if the value of 0.5 which is the desired case, is subtracted from the measure scores of male and female gender for a given list, then the remaining bias scores of each gender group are symmetric. Same holds for the exposure measure. Additionally, these adapted bias measures are expected to compute smoother and more realistic bias scores owing to the probability scores of the *perceived* gender groups instead of single labels which are too deterministic for annotating real datasets.

After the computation of adapted representation and exposure bias scores, the mean bias (MB) and mean absolute bias (MAB) of these measures can be further computed over a set of queries in the dataset to aggregate the bias results. MB score of STEM field computes a mean value over all the STEM queries' scores for the corresponding measure, whereas MAB computes a mean value over all the absolute value of the measure scores for the STEM queries. Note that MB shows towards which *perceived* gender the results are biased and MAB solves the limitation of MB if different queries have bias contributions with opposite signs and cancel each other out. Thus, MB and MAB measures are complementary for aggregating the results and interpreting those results in a proper way.

Please note that in the context of this chapter, the probability scores correspond to each *perceived* gender label, i.e. $prob_{G_m}(r_i)$ and $prob_{G_f}(r_i)$ for male and female respectively, of a given video is computed merely based on the narrators' *perceived* gender by using voice information of the narrators. For the details about the automated annotation procedure, please refer to 7.3.2. Since the probability scores are leveraged, there is no gender binary assumption throughout this chapter. As mentioned in Section 6.3.1, the proposed measures could easily be adapted to using the gender labels that are defined in a more fine-grained manner, e.g. probability scores.

7.2.2 Quantifying Bias

Using the measures of bias defined in Section 7.2.1, the *perceived* gender bias of the STEM and NON-STEM fields is first measured in *YVRP*s returned in response to the educational queries in various majors, and then a comparative evaluation is fulfilled.

- **Collecting *YVRP*s.** The same educational query list was used to crawl *YVRP*s, for the complete query list, please see Table 6.2. For each query, top-200 video results returned by YouTube UK in *incognito mode* was crawled by using the YouTube Data API v3 [1]. Note that the data collection process was done in a controlled environment such that the queries were sent to YouTube via the YouTube API by avoiding long time-lags. After the crawling, all the video results related to the majors in both STEM and NON-STEM fields, they were automatically assigned with probability scores of each *perceived* gender group using the voice information of the narrators.

- **Bias Evaluation.** The bias scores are computed for every *YVRP* with two adapted bias measures with four different cut-off values: $\mathcal{R}ep@n$ and $\mathcal{E}xp@n$ for $n = 3$, 10, 20, and 200, e.g. full list where $|r| = 200$. Then, the results are aggregated using the MB and MAB. Additionally, first the existence of bias for each field is examined, then the bias results of STEM and NON-STEM fields with different measures and cut-off values are compared. Subsequently, the impact of different cut-off values is investigated on bias scores of STEM and NON-STEM fields. Finally, the source of bias is investigated by comparing the top video search results with the full list, i.e. top-3, top-10 vs. top-200.

- **Statistical Analysis.** To identify whether the bias measured is not due to noise, a one-sample t-test is applied. Note that since the sample size is sufficiently large (> 30), according to the central limit theorem the sampling distribution is considered normal (Kwak & Kim, 2017). If this hypothesis is rejected, hence there is a significant difference and it can be claimed that the *YVRP*s of the evaluated field, STEM or NON-STEM is biased. The difference in bias measured across the two fields is further compared using a two-tailed *independent* t-test. In addition to the statistical significance, namely p-values, effect sizes are also reported using Cohen's d. Statistical significance helps to examine whether the findings show systematic bias or they are the result of noise, whereas effect sizes provide information about the magnitude of the differences. Thus, both p-values and effect sizes provide complementary information for the interpretation of the results.

  Apart from these, to investigate the effect of different cut-off values on bias results in the same field, STEM or NON-STEM, a two-tailed *paired* t-test is computed since the same query set is examined for different cut-off values and for two different annotation models. Moreover, Bonferroni correction is further applied (Sedgwick, 2012) for multiple hypothesis testing since there are 60 hypotheses in total in the context of cut-off value and annotation model analyses. Thus, without the Bonferroni correction, with the significance level,

[1] `https://developers.google.com/youtube/v3/docs`

$\alpha = .05$ and 60 hypotheses, the probability of identifying at least one significant result due to chance is around 0.95 which means that the results could be misleading. Hence, the Bonferroni correction is also applied for more reliable results in the scope of the cut-off value and annotation model analysis in Section 7.3.3. Note that for the significance level where $\alpha = .05$, and with the Bonferroni correction new $\alpha = .0008$. Thus, Bonferroni correction rejects the null hypothesis for each p-value (p_i) if $p_i <= .0008$ instead of .05. For the significance level where $\alpha = .01$, with the Bonferroni correction new $\alpha = .0002$.

7.3 Experimental Setup

In this section, first dataset information, then the annotation procedure, and lastly the *perceived* gender bias results will be provided based on the proposed method as described in Section 7.2. In addition to the existence of bias, source of *perceived* gender bias will be investigated as well.

7.3.1 Dataset

In this chapter, the main aim is to mimic the user scenario that has been fulfilled in Chapter 6 without a manual annotation. Unlike Chapter 6, *YVRP*s are annotated using an automated model and *perceived gender* label is automatically inferred from the voice of the narrators. For the detailed annotation procedure, please see 7.3.2.

The same educational query list was used to crawl the dataset. For the crawling, YouTube Data API v3 was utilised. Note that crawling with a Python implementation using the YouTube Data API v3 was fast enough that it does not create noticeable time lags between queries which could affect the bias analysis. Initially, several API keys were created for free. Then, using these API keys the location, i.e. *regionCode* in the API document, was set to the UK and *YVRP*s were crawled automatically. Yet, YouTube Data API v3 has some limitations as follows. First, one can crawl a limited number of *YVRP*s for each generated API key; the quota is based on the information crawled for each API request. Thus, it is important to crawl only sufficient information while using the YouTube Data API. For the detailed quota information, one can use the official page [2]. Second, the YouTube Data API v3 returns 50 *YVRP*s in total for each API request; this is the maximum number that could be retrieved using the official API per request. Therefore, one needs to find a workaround to crawl more than 50 *YVRP*s per query. For this purpose, the YouTube Data API v3 provides a field of *nextPageToken* that denotes a unique ID for the next page of the *YVRP*s of the current query. Then, this *nextPageToken* could be assigned to the *pageToken* field of the YouTube Data API v3 request while crawling the next page of the video search results of the same query.

Based on the aforementioned information, using the YouTube Data API v3 with the same educational query list in Chapter 6, in total 200 *YVRP*s were crawled for each query. Similarly, in the scope of this chapter, since personalised search might complicate the bias analysis, the analysis was designed in unpersonalised search settings, i.e. there was no user information included in the API requests. In addition to the insights in Chapter 6, the main reason of using automated crawling and annotation procedures in this chapter is to investigate the source of bias in search results as well. Hence, 200 *YVRP*s were obtained with the assumption that these search results could be the representative of the full video search results for the corresponding query that could be used to track the source of bias. Since crawling more *YVRP*s require huge processing time, especially in the annotation phase which uses deep learning-based automated models, 200 *YVRP*s were selected for the source of bias analysis. To crawl the 200 *YVRP*s for each query, four API requests were sent to the YouTube Data API v3 using the *nextPageToken* information. In this way, 200 results were obtained for all the educational queries, except the query of *Capital labour and power: Britain 1707-1939* in the NON-STEM major of *Politics*. Only for this query, YouTube Data API v3 returned 67 *YVRP*s in total and changing punctuations in the query etc. did not change the retrieved search results by YouTube. Please note that the educational query list was not modified on purpose in order to obtain comparative evaluation results with Chapter 6 and not to inject personal bias.

7.3.2 Annotation Procedure

[2] `https://developers.google.com/youtube/v3/determine_quota_cost`

The automated annotation procedure has two essential steps. First a video in a *YVRP* is annotated based on its relevancy with respect to the given query as *relevant*, *not-relevant*. If the video is *relevant* to the given query, then the video is annotated with a probability distribution of *male* and *female* genders based on the *perceived* gender of narrator(s) through voice information instead of labelling each video with a single *perceived* gender label as fulfilled in Section 6.4.1. Note that before these two main annotation steps, videos with the following properties were discarded.

– Short Videos: The videos that are shorter than 20 seconds.

– Unavailable Videos: The videos that has been removed by the user or from YouTube because of copyrights.

– Restricted Videos: The videos that require a user login to watch either because it is a private video or because of age restriction.

After discarding these videos, the first step of the annotation procedure was applied.

7.3.2.1 Relevance Annotation

After the crawling, to automatically detect the relevance label of a given video with respect to the corresponding query, a document similarity approach was implemented. *YVRP*s were converted into textual contents which were then used to measure document similarity. The main idea here is to utilise the dataset in Chapter 6 that had already been annotated with the relevance labels. The approach of measuring document similarity combines the following three models to encode documents: Term-frequency muliplied by inverse document-frequency (tf*idf) [3], Universal Sentence Encoders (USE) (Cer, Yang, Kong, Hua, Limtiaco, John, Constant, Guajardo-Cespedes, Yuan & Tar, 2018) [4], and Sentence-BERT (SBERT) (Reimers & Gurevych, 2019) using SentenceTransformers [5] Python framework. In addition to these, a jaccard similarity measure (Niwattanakul, Singthongchai, Naenudorn & Wanapu, 2013),

[3] `https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html`

[4] `https://tfhub.dev/google/universal-sentence-encoder/4`

[5] `https://github.com/UKPLab/sentence-transformers/blob/master/index.rst`

which is mainly based on the number of common words between two documents, was also implemented. Yet, the jaccard measure did not work well so it was discarded from the analysis.

After the implementation of these models, threshold values of each model were determined experimentally. For evaluating the capability of these models and tuning the threshold values on the textual contents of *YVRP*s, the dataset in Chapter 6 was used. Then the results of the aforementioned three models of tf*idf, USE, and SBERT were merged for measuring document similarity. To automatically label the relevancy of the *YVRP*s that were crawled with the data crawling procedure as described in Section 7.3.1, these steps were fulfilled:

(i) Using the *videoID*, title, description, and subtitles of each video in the *YVRP*s of were crawled.

(ii) Then, title, description and subtitles of each video were concatenated; each video was represented with this concatenated textual content.

(iii) A preprocessing phase of removing numbers, punctuation, making lowercase, and lemmatisation was applied.

(iv) The three models of tf*idf, USE, and SBERT were applied on the preprocessed textual contents of the *YVRP*s that were crawled and annotated in Chapter 6.

(v) Given the true relevancy label of a given document (video), experimentation was fulfilled with model representations of the given document and different threshold values for each model.

(vi) The threshold values of 0.1, 0.5 and 0.5 were determined for tf*idf, USE, and SBERT respectively. If the computed document similarity score is below a threshold, then the document is labelled as not-relevant for the corresponding model.

To automatically annotate a given video in the rest of the videos with a relevance label, three different models were deliberately used. After computing document similarity scores, these scores were reviewed as well. For the final decision about the relevancy of a given video, the document similarity scores of the three models were expected to be consistent. This means that a more conservative approach was taken, i.e. a given document/video is not-relevant if these three models agree, in order not to lose the relevant documents.

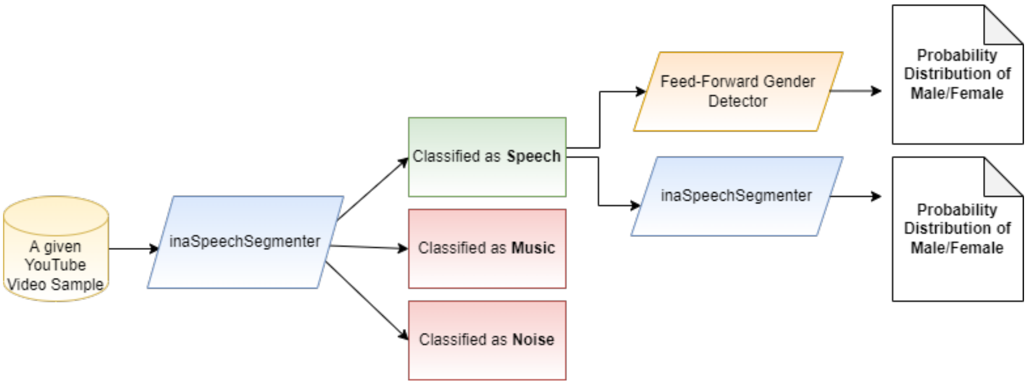

A given YouTube Video Sample
inaSpeechSegmenter
Classified as Speech
Classified as Music
Classified as Noise
Feed-Forward Gender Detector
inaSpeechSegmenter
Probability Distribution of Male/Female
Probability Distribution of Male/Female

Figure 7.1 Flow-chart of the *Perceived* Gender Annotation

Based on the reviewing of the document similarity scores obtained from these three models, it was observed that the combined approach worked sufficiently well. The combined approach has the capability of detecting non-English as well as out-of-context videos. For instance, YouTube returned some videos related to Python context managers for the query/course module of *Management in context* in *Public Relations* major of NON-STEM field and the combined approach detected labelled those videos as *not-relevant*. Additionally, the combined model detected some non-English videos for the query/course module of *Urban sociology* in *Sociology* major of NON-STEM field. Note that in the scope of this chapter, an additional relevance label of *N/A* was not used for non-English videos and videos without a narrator. Instead, non-English videos were labelled together with out-of-context videos as *not-relevant*. The videos with no narrator were handled during the *perceived* gender annotation in Section 7.3.2.2.

7.3.2.2 Perceived Gender Annotation

As the second phase of the annotation procedure, if a given video is relevant then it was annotated automatically with *perceived* gender information by computing its probability distribution of male and female gender labels based on the audio of the given video. This phase is composed of two main steps. First, the given audio was classified into the segments of speech, music, or noise using the model of inaSpeechSegmenter (Salmon & Vallet, 2014). Note that the audio segment of music or noise correspond to no narrator case, i.e. this refers to the label of *N/A*. Subsequently, the inaSpeechSegmenter and Feed-

Forward Gender Detector [6] models were used independently to detect *perceived* gender on the speech segments. Finally, male (female) ratio was computed by measuring the time of the audio segment that is annotated with male (female) *perceived* gender label divided by the time of the full speech audio segment. The *perceived* gender annotation procedure is displayed in Figure 7.1.

inaSpeechSegmenter proposes a gender detection processing pipeline which is composed of three main parts. The first submodule of a Speech/Music segmenter based on Convolutional Neural Networds (CNN) [7] is responsible for discarding music and empty segments. Subsequently, features corresponding to speech segments are extracted using a common extraction framework. A simple energy threshold is utilised to discard frames with low energy. Lastly, Gaussian Mixture Models (GMMs), i-vectors, i.e. compact vector representation of speech utterance, and CNN systems are then leveraged to classify the remaining speech segments into male and female excerpts. The inaSpeechSegmenter has been trained on the INA's Speaker Dictionary (Doukhan, Carrive, Vallet, Larcher & Meignier, 2018) which contains about 32000 excerpts of 1780 male (94 hours) and 494 female (27 hours) speakers. This audiovisual corpus was annotated with a semi-automatic labelling protocol. For more details about inaSpeechSegmenter, please refer to the original paper (Salmon & Vallet, 2014).

The second model of the feed-forward gender detector is a deep feed-forward neural network of five hidden layers, i.e. 0.3 dropout rate after each dense layer, was presented in this tutorial [8]. The dataset that was used in the second model is Mozilla's Common Voice Dataset [9] which is a corpus of speech data read by users on the Common Voice Website [10]. Before using the dataset, the dataset was balanced as the number of male samples equal to female samples, i.e. in total 67K samples with equal number of male/female samples, to prevent the model to favour one particular gender. In addition, for feature extraction this second model utilises Mel Spectogram [11] extraction technique to obtain a compact vector representation of length 128. For more details, please refer

[6] https://github.com/x4nth055/gender-recognition-by-voice

[7] https://en.wikipedia.org/wiki/Convolutional_neural_network

[8] https://www.thepythoncode.com/article/gender-recognition-by-voice-using-tensorflow-in-python

[9] https://www.kaggle.com/datasets/mozillaorg/common-voice

[10] https://commonvoice.mozilla.org/fr

[11] https://en.wikipedia.org/wiki/Mel_scale

to the tutorial and the implementation of the model.

For evaluating the capability of the aforementioned two automated models, the dataset of VoxCeleb [12] was used. VoxCeleb contains audio and video of short clip extracted from the interviews on YouTube. The dataset includes speakers from a wide range of ethnicities as well as ages. In terms of gender, the dataset is composed of 3682 male speakers (61%), and 2311 female (39%). Note that during the evaluation phase, only the audio clips were used since in the scope of this chapter, YouTube videos are annotated with probability distributions using only the audio information in a given video. Since the evaluation dataset is also from YouTube, model evaluation results for binary classification could be a good indicator in terms of the models' annotation capability. Based on the evaluation results shown in Table 7.2, it is observed that inaSpeechSegmenter outperforms Feed-Forward Gender Detector by a large margin – inaSpeechSegmenter gives aroung 20% higher F1-scores for the female class, while around 10% for the male class.

Note that the length of the educational videos crawled from YouTube might vary from couple of minutes to multiple hours. Hence, applying two deep learning-based models to get annotations might take a huge amount of time. To speed up the annotation procedure, rather than annotating the full video, a group of samples were selected from each video. Moreover, most of these educational videos start with a musical intro and end with a musical outro. Thus, in order to make the sampling more focused on the video content, the sampling was initiated 10 seconds after the beginning and terminated 10 seconds before the ending. Then, for the sampling of each video, i.e. after discarding both the first and last 10 seconds, first 10 seconds of each video minute was taken as a sample. Finally, on each sample the annotation steps as denoted in Figure 7.1 were executed. In this way, male/female probability distributions of each video were computed using the corresponding sample-level annotations of the aforementioned two models. Instead of *perceived* single labels, probability distributions are expected to provide smoother and more realistic bias results in comparison to the results in 6.4.2.

7.3.3 Results

[12] https://www.robots.ox.ac.uk/~vgg/data/voxceleb/

Table 7.2 Evaluation Results on VoxCeleb

Model	Gender	Precision	Recall	F1-score
Feed-Forward Gender Detector	Female	0.79	0.79	0.79
	Male	0.87	0.87	0.87
inaSpeechSegmenter	Female	0.92	0.94	0.93
	Male	0.96	0.95	0.96

Table 7.3 Feed-Forward Gender Detector: *Perceived* gender bias in YouTube for the top-20 relevant results where $n = 3, 10, 20$, p-values of a two-tailed independent t-test computed between the **MB and MAB** scores of STEM and NON-STEM fields

		$\mathcal{R}ep$@3	$\mathcal{R}ep$@10	$\mathcal{R}ep$@20	$\mathcal{E}xp$@3	$\mathcal{E}xp$@10	$\mathcal{E}xp$@20
MB	STEM	0.2881***	0.2651***	0.2352***	0.3059***	0.2777***	0.2525***
	NON-STEM	0.2209***	0.2225***	0.1868***	0.2424***	0.2324***	0.2050***
	p-value	.53	.51	.41	.56	.52	.43
	effect size d	0.129	0.133	0.169	0.119	0.131	0.159
MAB	STEM	0.5018***	0.3189***	0.2842***	0.5227***	0.3426***	0.3005***
	NON-STEM	0.4893***	0.3667***	0.3139***	0.5026***	0.4028***	0.3425***
	p-value	.84	.25	.45	.75	.16	.29
	effect size d	0.041	-0.233	-0.154	0.065	-0.286	-0.215

First, it is determined if YouTube returns biased results in terms of the narrator's *perceived gender* annotated with probability scores for STEM and NON-STEM fields separately, and if so, whether YouTube's top-200 relevant search results suffer from the same level of bias, such that the difference between STEM and NON-STEM queries/course modules is not statistically significant. Further the source of bias is tracked, if bias exists then it is investigated whether the bias comes from the input data or the ranking algorithm itself. Note that computing probability scores for each *perceived* gender group was fulfilled by using two different automated models, namely *Feed-Forward Gender Detector* and *inaSpeechSegmenter* models. Thus, *perceived* gender bias is measured for these two models separately. For more information about the models, please refer to Section 7.3.2.2.

7.3.3.1 Feed-Forward Gender Detector Model Results

In Table 7.3, the *perceived* gender bias is displayed using the adapted measures presented in Section 7.2.1, namely $\mathcal{R}ep_{prob}@n$ and $\mathcal{E}xp_{prob}@n$ for different cut-

Table 7.4 Feed-Forward Gender Detector: *Perceived* gender bias in YouTube for the top-20 relevant results where $n = 3$, 10, 20, p-values of a two-tailed paired t-test computed between the **MB** scores of STEM and NON-STEM fields

		STEM	NON-STEM		STEM	NON-STEM		STEM	NON-STEM
	$\mathcal{R}ep$@3	0.2881***	0.2209***	$\mathcal{R}ep$@10	0.2651***	0.2225***	$\mathcal{R}ep$@3	0.2881***	0.2209***
	$\mathcal{R}ep$@10	0.2651***	0.2225***	$\mathcal{R}ep$@20	0.2352***	0.1868***	$\mathcal{R}ep$@20	0.2352***	0.1868***
MB	p-value	.72	.97		.24	.12		.42	.57
	effect size d	0.057	−0.004		0.117	0.104		0.134	0.077
	$\mathcal{E}xp$@3	0.3059***	0.2424***	$\mathcal{E}xp$@10	0.2777***	0.2324***	$\mathcal{E}xp$@3	0.3059***	0.2424***
	$\mathcal{E}xp$@10	0.2777***	0.2324***	$\mathcal{E}xp$@20	0.2525***	0.2050***	$\mathcal{E}xp$@20	0.2525***	0.2050***
	p-value	.58	.80		.18	.13		.35	.45
	effect size d	0.066	0.021		0.093	0.075		0.129	0.083

Table 7.5 Feed-Forward Gender Detector: *Perceived* gender bias in YouTube for the top-20 relevant results where $n = 3$, 10, 20, p-values of a two-tailed paired t-test computed between the **MAB** scores of STEM and NON-STEM fields

		STEM	NON-STEM		STEM	NON-STEM		STEM	NON-STEM
	$\mathcal{R}ep$@3	**0.5018*** **	0.4893***	$\mathcal{R}ep$@10	0.3189***	0.3677***	$\mathcal{R}ep$@3	**0.5018*** **	**0.4893*** **
	$\mathcal{R}ep$@10	**0.3189*** **	0.3677***	$\mathcal{R}ep$@20	0.2842***	0.3139***	$\mathcal{R}ep$@20	**0.2842*** **	**0.3139*** **
MAB	p-value	**.0002**	.0051		.086	.0083		**< .0001**	**.0002**
	effect size d	**0.709**	0.461		0.180	0.257		**0.866**	**0.681**
	$\mathcal{E}xp$@3	**0.5227*** **	0.5026***	$\mathcal{E}xp$@10	0.3426***	**0.4028*** **	$\mathcal{E}xp$@3	**0.5227*** **	**0.5026*** **
	$\mathcal{E}xp$@10	**0.3426*** **	0.4028***	$\mathcal{E}xp$@20	0.3005***	**0.3425*** **	$\mathcal{E}xp$@20	**0.3005*** **	**0.3425*** **
	p-value	**< .0001**	.0026		.0193	**.0002**		**< .0001**	**.0001**
	effect size d	**0.674**	0.382		0.212	**0.291**		**0.857**	**0.621**

Table 7.6 Feed-Forward Gender Detector: *Perceived* gender bias in YouTube for all the relevant results crawled, i.e. $|r| = 200$ where $n = 3$, 10, 20, 200, p-values of a two-tailed paired t-test computed between the **MB and MAB** scores of STEM and NON-STEM fields using the measure of $\mathcal{R}ep@n$.

		STEM	NON-STEM		STEM	NON-STEM		STEM	NON-STEM
MB	$\mathcal{R}ep$@3	0.2881***	0.2209***	$\mathcal{R}ep$@10	0.2651***	0.2225***	$\mathcal{R}ep$@20	0.2352***	0.1868***
	$\mathcal{R}ep$@200	0.1879***	0.1564***	$\mathcal{R}ep$@200	0.1879***	0.1564***	$\mathcal{R}ep$@200	0.1879***	0.1564***
	p-value	.16	.32		.0405	.0353		.0905	.14
	effect size d	0.250	0.155		0.301	0.211		0.194	0.105
MAB	$\mathcal{R}ep$@3	**0.5018*** **	**0.4893*** **	$\mathcal{R}ep$@10	0.3189***	**0.3677*** **	$\mathcal{R}ep$@20	0.2842***	**0.3139*** **
	$\mathcal{R}ep$@200	**0.2750*** **	**0.2457*** **	$\mathcal{R}ep$@200	0.2750***	**0.2457*** **	$\mathcal{R}ep$@200	0.2750***	**0.2457*** **
	p-value	**< .0001**	**< .0001**		.0781	**< .0001**		.71	**.0001**
	effect size d	**0.954**	**0.991**		0.253	**0.627**		0.056	**0.366**

off values of $n = 3$, 10, 20. All MB and MAB scores are positive for both bias measures; the one-sample t-test computed on MBs and MABs are statistically significant for the measures where p-value < .001 denoted as ***. The two-tailed paired t-test computed on MB and MABs to compare the difference in

Table 7.7 Feed-Forward Gender Detector: *Perceived* gender bias for specific majors of STEM and NON-STEM fields in YouTube for the top-20 relevant results - red denotes bias towards male while blue towards female

	Biology	*Chemistry*	*CS*	*Maths*	*Physics*	*Eng. Lan. Lit.*	*Politics*	*Psychology*	*Pub. Rel.*	*Sociology*
$\mathcal{R}ep$@3	**0.0821**	0.0808	0.3952	**0.4561**	0.4264	0.0721	**0.3441**	0.4354	0.3471	**-0.0945**
$\mathcal{R}ep$@10	**0.0135**	0.2112	0.2852	0.3868	**0.4288**	0.0755	**0.4511**	0.2978	0.2466	**0.0418**
$\mathcal{R}ep$@20	**0.0308**	0.1568	0.2580	0.2722	**0.4580**	0.0087	**0.4414**	0.2352	0.2385	**0.0103**
$\mathcal{E}xp$@3	0.1328	**0.0785**	0.3729	**0.5121**	0.4333	0.1800	0.3943	**0.4950**	0.2364	**-0.0935**
$\mathcal{E}xp$@10	**0.0488**	0.1872	0.2904	**0.4344**	0.4275	0.1381	**0.4502**	0.3552	0.2132	**0.0053**
$\mathcal{E}xp$@20	**0.0472**	0.1599	0.2667	0.3371	**0.4513**	0.0704	**0.4447**	0.2944	0.2221	**-0.0065**

bias between the STEM and NON-STEM fields, the results indicate that their differences are statistically not significant on the bias measures of $\mathcal{R}ep_{prob}@n$ and $\mathcal{E}xp_{prob}@n$ where $n = 3, 10, 20$. Some effect sizes that correspond to the difference of bias using MAB scores are negative which indicates that the MAB score of the *perceived* gender group of female is higher than male, albeit statistically not significant.

Regarding the impact of different cut-off values, in Table 7.4 using MB scores, it is observed that, both STEM and NON-STEM fields show similar scores for different cut-off values on both $\mathcal{R}ep_{prob}@n$ and $\mathcal{E}xp_{prob}@n$ measures, i.e. the two-tailed paired t-test computed on MB scores are statistically not significant. On the other hand, in Table 7.5 using MAB scores, it is observed that cut-off values might affect the *perceived* gender bias in STEM and NON-STEM fields. The two-tailed paired t-test computed on MABs of STEM field is statistically significant for both measures of $\mathcal{R}ep_{prob}@n$ and $\mathcal{E}xp_{prob}@n$ between the following cut-off values, $n = 3$ vs. $n = 10$ and $n = 3$ vs. $n = 20$ with difference confidence levels. For the measure of $\mathcal{R}ep_{prob}@n$ between $n = 3$ vs. $n = 10$, p-value which corresponds to the significance level of $\alpha = .01$, and $n = 3$ vs. $n = 20$, p-value = .0002 corresponds to the significance level of $\alpha = .01$ and p-value = 0.000001 corresponds to the significance level of $\alpha = .005$ respectively after Bonferroni correction was applied. For the measure of $\mathcal{E}xp_{prob}@n$ between $n = 3$ vs. $n = 10$ and $n = 3$ vs. $n = 20$, p-value = .00005 which corresponds to the significance level of $\alpha = .01$ and p-value = .000003 which corresponds to the significance level of $\alpha = .0001$ respectively with Bonferroni correction. The two-tailed paired t-test computed on MABs of NON-STEM field is statistically significant for both measures of $\mathcal{R}ep_{prob}@n$ and $\mathcal{E}xp_{prob}@n$ between different cut-off values. For the measure of $\mathcal{R}ep_{prob}@n$ the difference is statistically significant only between $n = 3$ vs. $n = 20$, p-value = .0002 which corresponds to the significance level of $\alpha = .01$. For the measure of $\mathcal{E}xp_{prob}@n$ between $n = 10$ vs. $n = 20$ and $n = 3$ vs. $n = 20$, p-value = .0002 which corresponds to the significance level of $\alpha = .05$ and p-value = .0001 which corresponds to the significance level of $\alpha = .01$ respectively with Bonferroni correction.

For tracking the source of bias, only the bias scores from the measure of $\mathcal{R}ep_{prob}@n$ are used since the strong sense of rank information, i.e. $\mathcal{E}xp_{prob}@n$, is not meaningful where $n = 200$ from the user's perspective. In Table 7.6, the two-tailed paired t-test computed on MBs of STEM and NON-STEM fields are statistically not significant. Unlike, the two-tailed paired t-test computed on MABs of STEM field is statistically significant only between $n = 3$ vs. $n = 200$, p-value = .000004 which corresponds to the significance level of $\alpha = .0005$ with Bonferroni correction. On the other hand, the two-tailed paired t-test computed on MABs of NON-STEM field is statistically significant between $n = 3$ vs. $n = 200$, $n = 10$ vs. $n = 200$, and $n = 20$ vs. $n = 200$, p-value = .000003 which corresponds to the significance level of $\alpha = .0005$, p-value = .000008 to the $\alpha = .0005$, and p-value = .0001 to the $\alpha = .01$ respectively with Bonferroni correction.

In Table 7.7, the bias scores for each major in STEM and NON-STEM fields are displayed using the measures $\mathcal{R}ep_{prob}@n$ and $\mathcal{E}xp_{prob}@n$ for different cut-off values. Note that the highest/lowest bias scores are denoted as highlighted. In Figure 7.2a, the overall *perceived* gender bias scores are compared for STEM and NON-STEM fields on the MB scores of $\Delta_{\mathcal{E}xp_{prob}@10}$ using the Feed-Forward Gender Detector model. It is observed that STEM is more biased towards the male than NON-STEM field. In Figure 7.3 (a) and (b), the impact of different cut-off values is displayed on bias scores of $\Delta_{\mathcal{R}ep_{prob}@n}$ where $n = 3$ vs. $n = 10$ for STEM and NON-STEM fields. Similarly, in Figure 7.4 (a) and (b), the measure of $\Delta_{\mathcal{E}xp_{prob}@n}$ is used for the same purpose where $n = 3$ vs. $n = 10$. Note that both of these figures visualise the effect of different cut-off values on the *perceived* gender bias using the Feed-Forward Gender Detector model.

7.3.3.2 inaSpeechSegmenter Model Results

In Table 7.8, the *perceived* gender bias is displayed using the adapted measures for different cut-off values of $n = 3, 10, 20$. All MB and MAB scores are positive for both bias measures; the one-sample t-test computed on MBs

Table 7.8 inaSpeechSegmenter: *Perceived* gender bias in YouTube for the top-20 relevant results where $n = 3, 10, 20$, p-values of a two-tailed independent t-test computed between the **MB and MAB** scores of STEM and NON-STEM fields

		$\mathcal{R}ep$@3	$\mathcal{R}ep$@10	$\mathcal{R}ep$@20	$\mathcal{E}xp$@3	$\mathcal{E}xp$@10	$\mathcal{E}xp$@20
MB	STEM	0.4669***	0.4033***	0.3850***	0.4738***	0.4254***	0.4049***
	NON-STEM	0.1818***	0.2303***	0.2323***	0.1869***	0.2228***	0.2288***
	p-value	.0109	.0143	.0195	.0122	.0073	.0084
	effect size d	0.525	0.505	0.480	0.516	0.555	0.545
MAB	STEM	0.5906***	0.4189***	0.4039***	0.5968***	0.4469***	0.4194***
	NON-STEM	0.5350***	0.3849***	0.3454***	0.5462***	0.4181***	0.3630***
	p-value	.39	.51	.24	.46	.57	.25
	effect size d	0.173	0.135	0.237	0.152	0.117	0.233

Table 7.9 inaSpeechSegmenter: *Perceived* gender bias in YouTube for the top-20 relevant results where $n = 3, 10, 20$, p-values of a two-tailed paired t-test computed between the **MB** scores of STEM and NON-STEM fields

		STEM	NON-STEM		STEM	NON-STEM		STEM	NON-STEM
MB	$\mathcal{R}ep$@3	0.4669***	0.1818***	$\mathcal{R}ep$@10	0.4033***	0.2303***	$\mathcal{R}ep$@3	0.4669***	0.1818***
	$\mathcal{R}ep$@10	0.4033***	0.2303***	$\mathcal{R}ep$@20	0.3850***	0.2323***	$\mathcal{R}ep$@20	0.3850***	0.2323***
	p-value	.31	.42		.42	.94		.22	.48
	effect size d	0.162	−0.095		0.069	−0.005		0.209	−0.103
	$\mathcal{E}xp$@3	0.4738***	0.1869***	$\mathcal{E}xp$@10	0.4254***	0.2228***	$\mathcal{E}xp$@3	0.4738***	0.1869***
	$\mathcal{E}xp$@10	0.4254***	0.2228***	$\mathcal{E}xp$@20	0.4049***	0.2288***	$\mathcal{E}xp$@20	0.4049***	0.2288***
	p-value	.31	.43		.24	.78		.22	.47
	effect size d	0.119	−0.068		0.076	−0.015		0.173	−0.083

Table 7.10 inaSpeechSegmenter: *Perceived* gender bias in YouTube for the top-20 relevant results where $n = 3, 10, 20$, p-values of a two-tailed paired t-test computed between the **MAB** scores of STEM and NON-STEM fields

		STEM	NON-STEM		STEM	NON-STEM		STEM	NON-STEM
MAB	$\mathcal{R}ep$@3	0.5906***	0.5350***	$\mathcal{R}ep$@10	0.4189***	0.3849***	$\mathcal{R}ep$@3	**0.5906*****	**0.5350*****
	$\mathcal{R}ep$@10	0.4189***	0.3849***	$\mathcal{R}ep$@20	0.4039***	0.3454***	$\mathcal{R}ep$@20	**0.4039*****	**0.3454*****
	p-value	.0010	.0025		.42	.11		**.0005**	**.0003**
	effect size d	0.596	0.519		0.063	0.152		**0.656**	**0.658**
	$\mathcal{E}xp$@3	**0.5968*****	0.5462***	$\mathcal{E}xp$@10	0.4469***	0.4181***	$\mathcal{E}xp$@3	**0.5968*****	**0.5462*****
	$\mathcal{E}xp$@10	**0.4469*****	0.4181***	$\mathcal{E}xp$@20	0.4194***	0.3630***	$\mathcal{E}xp$@20	**0.4194*****	**0.3630*****
	p-value	**.0005**	.0013		.11	.0048		**.0002**	**.0002**
	effect size d	**0.497**	0.448		0.114	0.222		**0.601**	**0.636**

and MABs are statistically significant for the measures where p-value $< .001$ denoted as ***. The two-tailed paired t-test computed on MB and MABs to compare the difference in bias between the STEM and NON-STEM fields, the results indicate that their differences are statistically not significant on the bias measures of $\mathcal{R}ep_{prob}@n$ and $\mathcal{E}xp_{prob}@n$ where $n = 3, 10, 20$. Since

Table 7.11 inaSpeechSegmenter: *Perceived* gender bias in YouTube for all the relevant results crawled, i.e. $|r| = 200$ where $n = 3, 10, 20, 200$, p-values of a two-tailed paired t-test computed between the **MB and MAB** scores of STEM and NON-STEM fields using the measure of $\mathcal{R}ep@n$.

		STEM	NON-STEM		STEM	NON-STEM		STEM	NON-STEM
MB	$\mathcal{R}ep@3$	0.4669***	0.1818***	$\mathcal{R}ep@10$	0.4033***	0.2303***	$\mathcal{R}ep@20$	0.3850***	0.2323***
	$\mathcal{R}ep@200$	0.1816***	0.2194***	$\mathcal{R}ep@200$	0.1816***	0.2194***	$\mathcal{R}ep@200$	0.1816***	0.2194***
	p-value	.15	.62		.27	.74		.49	.53
	effect size d	0.236	-0.080		0.119	0.030		0.062	0.038
MAB	$\mathcal{R}ep@3$	**0.5906*** **	**0.5350*** **	$\mathcal{R}ep@10$	0.4189***	0.3849***	$\mathcal{R}ep@20$	0.4039***	0.3454***
	$\mathcal{R}ep@200$	**0.4001*** **	**0.3108*** **	$\mathcal{R}ep@200$	0.4001***	0.3108***	$\mathcal{R}ep@200$	0.4001***	0.3108***
	p-value	**.0002**	**< .0001**		.52	.0124		.89	.08
	effect size d	**0.522**	**0.830**		0.057	0.310		0.012	0.145

Table 7.12 inaSpeechSegmenter: *Perceived* gender bias for specific majors of STEM and NON-STEM fields in YouTube for the top-20 relevant results - red denotes bias towards male while blue towards female

	Biology	*Chemistry*	*CS*	*Maths*	*Physics*	*Eng. Lan. Lit.*	*Politics*	*Psychology*	*Pub. Rel.*	*Sociology*
$\mathcal{R}ep@3$	**0.2419**	0.3889	0.4883	**0.6779**	0.5372	0.1076	**0.4098**	0.4098	0.1443	**-0.2027**
$\mathcal{R}ep@10$	**0.1519**	0.2940	0.3712	**0.6105**	0.5890	0.0763	**0.5319**	0.2229	-0.0152	**0.0763**
$\mathcal{R}ep@20$	**0.1553**	0.2414	0.3923	0.5122	**0.6238**	0.0642	**0.5366**	0.3001	0.2272	**0.0335**
$\mathcal{E}xp@3$	**0.2928**	0.3663	0.4895	**0.7011**	0.5192	0.2521	**0.4451**	0.4398	0.0269	**-0.2295**
$\mathcal{E}xp@10$	**0.1922**	0.3196	0.4012	**0.6457**	0.5682	0.1656	**0.5091**	0.3740	0.1458	**-0.0806**
$\mathcal{E}xp@20$	**0.1787**	0.2757	0.4039	0.5644	**0.6015**	0.1251	**0.5228**	0.3406	0.1803	**-0.0248**

Table 7.13 Annotation Model Analysis: *Perceived* gender bias in YouTube for the top-20 relevant results where $n = 3, 10, 20$, p-values of a two-tailed paired t-test computed between the **MB and MAB** scores of **STEM** field

		$\mathcal{R}ep@3$	$\mathcal{R}ep@10$	$\mathcal{R}ep@20$	$\mathcal{E}xp@3$	$\mathcal{E}xp@10$	$\mathcal{E}xp@20$
MB	Feed-Forward Gender Detector	**0.2881*** **	**0.2651*** **	**0.2352*** **	**0.3059*** **	**0.2777*** **	**0.2525*** **
	inaSpeechSegmenter	**0.4669*** **	**0.4033*** **	**0.3850*** **	**0.4738*** **	**0.4254*** **	**0.4049*** **
	p-value	**.0001**	**< .0001**	**< .0001**	**.0002**	**< .0001**	**< .0001**
	effect size d	**-0.358**	**-0.516**	**-0.590**	**-0.326**	**-0.525**	**-0.601**
MAB	Feed-Forward Gender Detector	0.5018***	0.3189***	**0.2842*** **	0.5227***	0.3426***	**0.3005*** **
	inaSpeechSegmenter	0.5906***	0.4189***	**0.4039*** **	0.5968***	0.4469***	**0.4194*** **
	p-value	.0261	.0006	**< .0001**	.0335	.0008	**.0001**
	effect size d	-0.281	-0.449	**-0.569**	-0.224	-0.453	**-0.560**

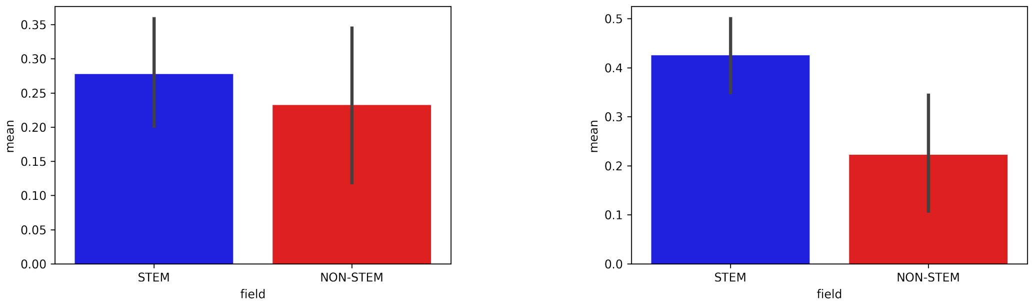
Bonferroni correction was applied, the results are not statistically significant.

Regarding the impact of different cut-off values, in Table 7.9 using MB scores, it is observed that, both STEM and NON-STEM fields show similar scores for

Table 7.14 Annotation Model Analysis: *Perceived* gender bias in YouTube for the top-20 relevant results where n = 3, 10, 20, p-values of a two-tailed paired t-test computed between the **MB and MAB** scores of **NON-STEM** field

		$\mathcal{R}ep@3$	$\mathcal{R}ep@10$	$\mathcal{R}ep@20$	$\mathcal{E}xp@3$	$\mathcal{E}xp@10$	$\mathcal{E}xp@20$
MB	Feed-Forward Gender Detector	0.2209***	0.2225***	0.1868***	0.2424***	0.2324***	0.2050***
	inaSpeechSegmenter	0.1818***	0.2303***	0.2323***	0.1869***	0.2228***	0.2288***
	p-value	.37	.79	.05	.25	.76	.36
	effect size d	0.069	-0.020	-0.132	0.097	0.023	-0.066
MAB	Feed-Forward Gender Detector	0.4893***	0.3677***	0.3139***	0.5026***	0.4028***	0.3425***
	inaSpeechSegmenter	0.5350***	0.3849***	0.3454***	0.5462***	0.4181***	0.3630***
	p-value	.24	.51	.13	.30	.55	.36
	effect size d	-0.148	-0.072	-0.136	-0.139	-0.066	-0.090

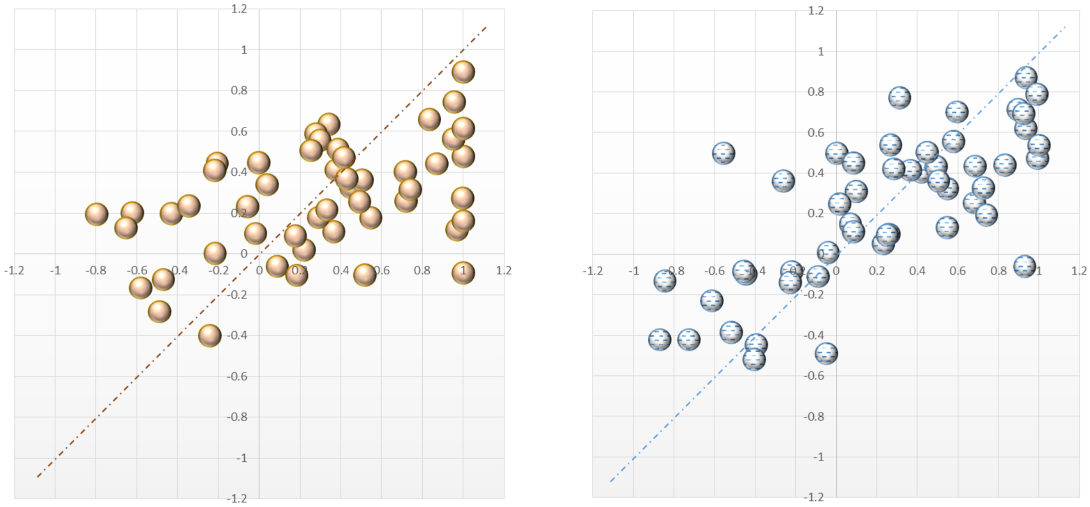
Figure 7.2 MB scores of $\Delta_{\mathcal{E}xp_{prob}@10}$ measured on *perceived* gender probability scores of STEM and NON-STEM fields.



(a) Feed-Forward Gender Detector

(b) inaSpeechSegmenter

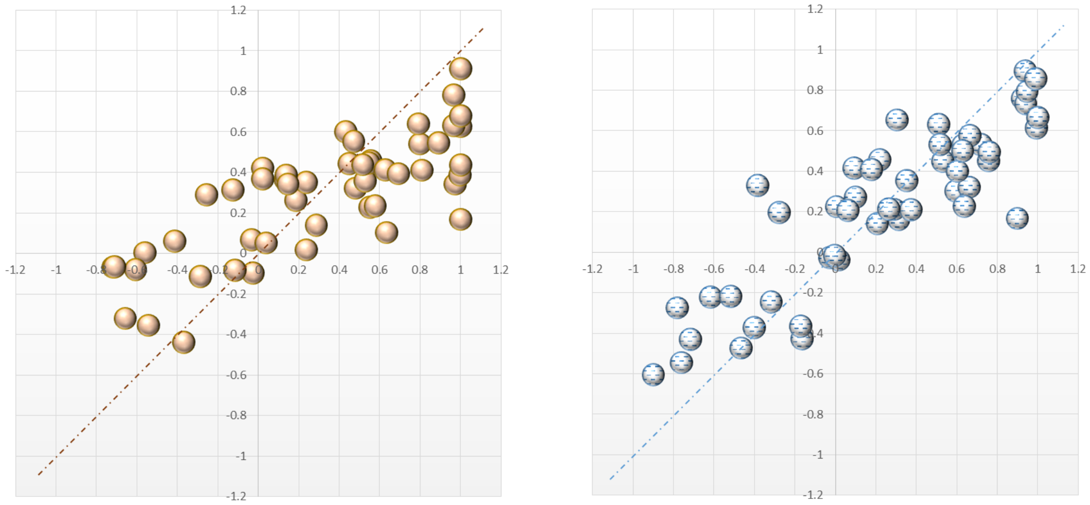
Figure 7.3 $\Delta_{\mathcal{R}ep_{prob}@n}$ measured on *perceived* gender probability scores of the **Feed-Forward Gender Detector**, where x-axis denotes $n = 3$ and y-axis $n = 10$.



(a) STEM field

(b) NON-STEM field

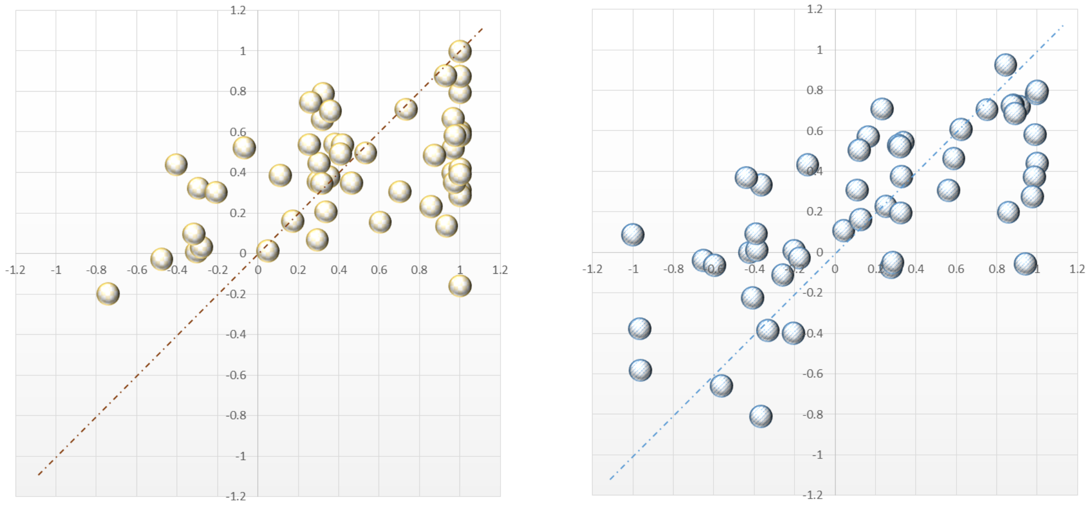
Figure 7.4 $\Delta_{\mathcal{E}xp_{prob}@n}$ measured on *perceived* gender probability scores of the **Feed-Forward Gender Detector**, where x-axis denotes $n = 3$ and y-axis $n = 10$.



(a) STEM field

(b) NON-STEM field

Figure 7.5 $\Delta_{\mathcal{R}ep_{prob}@n}$ measured on *perceived* gender probability scores of the **inaSpeechSegmenter**, where x-axis denotes $n = 3$ and y-axis $n = 10$.

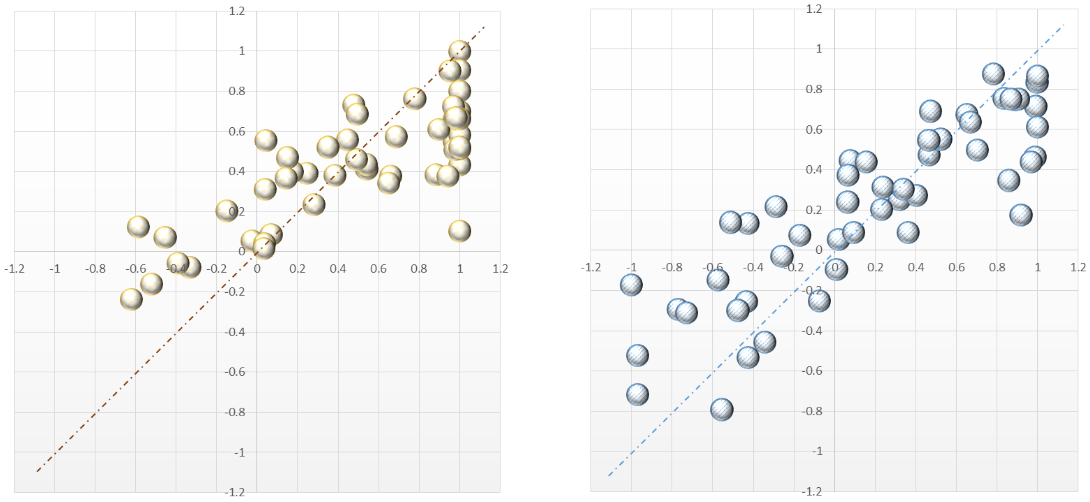


(a) STEM field

(b) NON-STEM field

different cut-off values on both $\mathcal{R}ep_{prob}@n$ and $\mathcal{E}xp_{prob}@n$ measures, i.e. the two-tailed paired t-test computed on MB scores are statistically not significant. Some effect sizes that correspond to the difference of bias using MB scores are negative, i.e. the MB score of the *perceived* gender group of female is higher than male, albeit statistically not significant. On the other hand, in Table 7.10 using MAB scores, it is observed that cut-off values might affect the *perceived* gender bias in STEM and NON-STEM fields. The two-tailed paired t-test computed on MABs of STEM field is statistically significant for both measures

Figure 7.6 $\Delta_{\mathcal{E}xp_{prob}@n}$ measured on *perceived* gender probability scores of the **inaSpeechSegmenter**, where x-axis denotes $n = 3$ and y-axis $n = 10$.



(a) STEM field **(b)** NON-STEM field

of $\mathcal{R}ep_{prob}@n$ and $\mathcal{E}xp_{prob}@n$. For the measure of $\mathcal{R}ep_{prob}@n$, between only the cut-off values, $n = 3$ vs. $n = 20$, p-value = .0005 which corresponds to the significance level of $\alpha = .05$ with Bonferroni correction. For the measure of $\mathcal{E}xp_{prob}@n$, between the following cut-off values, $n = 3$ vs. $n = 10$ and $n = 3$ vs. $n = 20$ with difference confidence levels, p-value = .0005 which corresponds to the significance level of $\alpha = .05$ and p-value = .0002 which corresponds to the significance level of $\alpha = .05$ respectively with Bonferroni correction.

Similarly, the two-tailed paired t-test computed on MABs of NON-STEM field is statistically significant for both measures of $\mathcal{R}ep_{prob}@n$ and $\mathcal{E}xp_{prob}@n$ between different cut-off values. For the measure of $\mathcal{R}ep_{prob}@n$ the difference is statistically significant only between $n = 3$ vs. $n = 20$, p-value = .0003 which corresponds to the significance level of $\alpha = .05$ with Bonferroni correction. For the measure of $\mathcal{E}xp_{prob}@n$ the difference is statistically significant only between $n = 3$ vs. $n = 20$, p-value = .0002 which corresponds to the significance level of $\alpha = .05$ respectively with Bonferroni correction.

For tracking the source of bias, again only the bias scores from the measure of $\mathcal{R}ep_{prob}@n$ are used. In Table 7.11, the two-tailed paired t-test computed on MBs of STEM and NON-STEM fields are statistically not significant. Unlike, the two-tailed paired t-test computed on MABs of STEM field is statistically significant only between $n = 3$ vs. $n = 200$, p-value = .0002 which corresponds to the significance level of $\alpha = .05$ with Bonferroni correction. Likewise, the two-tailed paired t-test computed on MABs of NON-STEM field is statistically significant only between $n = 3$ vs. $n = 200$, p-value = .00008 corresponds to the $\alpha = .005$ with Bonferroni correction.

In Table 7.12, the bias scores for each major in STEM and NON-STEM fields are displayed using the measures $\mathcal{R}ep_{prob}@n$ and $\mathcal{E}xp_{prob}@n$ for different cut-off values. Note that the highest/lowest bias scores are denoted as highlighted. In Figure 7.2b, the overall *perceived* gender bias scores are compared for STEM and NON-STEM fields on the MB scores of $\Delta_{\mathcal{E}xp_{prob}@10}$ using the inaSpeechSegmenter model. Table 7.13 displays comparison of the bias scores of the two annotation models only for the STEM field. The two-tailed paired t-test computed on MBs of the STEM field, the bias differences are statistically significant for both measures with different confidence values. On the measure of $\mathcal{R}ep_{prob}@n$, p-value = .0001 which corresponds to the significance level of $\alpha = .001$ and p-value = .000007 which corresponds to the significance level of $\alpha = .0005$ and p-value = .00000003 which corresponds to the significance level of $\alpha = .0001$ respectively for $n = 3$, $n = 10$, and $n = 20$ with Bonferroni correction. On the measure of $\mathcal{E}xp_{prob}@n$, p-value = .00002 which corresponds to the significance level of $\alpha = .001$, p-value = .000004 which corresponds to the significance level of $\alpha = .0005$ and p-value = .0000002 which corresponds to the significance level of $\alpha = .0001$ respectively for $n = 3$, $n = 10$, and $n = 20$ with Bonferroni correction. STEM field is biased using both models, towards the male gender (all MBs are positive) and inaSpeechSegmenter provides higher bias scores for the STEM field. Unlike, the two-tailed paired t-test computed on MABs of the STEM field is statistically significant for both measures only for $n = 20$ with the confidence levels of $\alpha = .0001$ and $\alpha = .0005$ for$\mathcal{R}ep_{prob}@n$ and $\mathcal{E}xp_{prob}@n$ respectively. For the NON-STEM field, NON-STEM fields show similar level of bias irrespective of the annotation model – the two-tailed paired t-test computed on MBs/MABs of the NON-STEM field is statistically not significant. This is verified across two measures with different cut-off values.

Figure 7.2 displays the comparison of the bias scores in STEM and NON-STEM fields for the Feed-Forward Gender Detector and inaSpeechSegmenter models. The error bars show the standard error on the scores of the corresponding field.

In Figure 7.5 (a) and (b), the impact of different cut-off values is displayed on bias scores of $\Delta_{\mathcal{R}ep_{prob}@n}$ where $n = 3$ vs. $n = 10$ for STEM and NON-STEM fields. Similarly, in Figure 7.6 (a) and (b), the measure of $\Delta_{\mathcal{E}xp_{prob}@n}$ is used for the same purpose where $n = 3$ vs. $n = 10$. Note that both of these figures visualise the effect of different cut-off values on the *perceived* gender bias using the inaSpeechSegmenter model.

7.4 Concluding Discussion

Initially, it is verified if the *YVRP*s are biased using the adapted measures of $\mathcal{R}ep_{prob}@n$ and $\mathcal{E}xp_{prob}@n$ (**RQ1**). If so, then it is investigated if the *YVRP*s suffer from the different magnitude of bias (**RQ2**) by examining if the difference between the bias scores of the *YVRP*s of STEM and NON-STEM fields are statistically significant. In Table 7.3 and Table 7.8 using Feed-Forward Gender Detector and inaSpeechSegmenter models respectively, regarding the **RQ1** the *YVRP*s of STEM and NON-STEM fields are both biased – the one-sample t-test applied on MB/MAB scores to check the existence of bias is statistically significant with p-value $< .0001$ as mentioned in Section 7.2.2. These findings suggest that both STEM and non-STEM fields are biased towards male (all MB scores are positive). On the basis of MAB scores, it is observed that both STEM and NON-STEM exhibit an absolute bias. Regarding the **RQ2**, STEM and NON-STEM fields show similar levels of bias – the two tailed *independent* t-tests applied on MB/MAB scores in Table 7.3 and Table 7.8. The differences of bias are statistically not significant irrespective of the measure, cut-off value **RQ3** and the automated model (Feed-Forward Gender Detector or inaSpeechSegmenter). With respect to the **RQ3**, it is also examined whether different cut-off values affect the existence of bias, the results are in these tables indicate that both STEM and NON-STEM fields are biased regardless of the cut-off values. Note that both groups of measures in Table 7.3 and Table 7.8 show consistent results. Nonetheless, Table 7.8 shows higher MB scores for the STEM field which implies that the inaSpeechSegmenter model produces more probability scores towards the male.

Regarding the **RQ4**, it is investigated whether different cut-off values change the magnitude of bias – the two tailed *paired* t-tests applied on MB scores are statistically not significant regardless of the model, see Table 7.4 and Table 7.9. Unlike the MB scores, the two tailed *paired* t-tests applied on MAB scores show some statistically significant results both for the Feed-Forward Gender Detector and inaSpeechSegmenter models. For the first model, in Table 7.5, the bias differences of $\mathcal{R}ep_{prob}@3$ and $\mathcal{R}ep_{prob}@10$ and $\mathcal{R}ep_{prob}@3$

and $\mathcal{R}ep_{prob}@20$ are statistically significant for the STEM field and the latter shows a higher difference (effect size of 0.866). This indicates that the STEM field, using the measure of $\mathcal{R}ep_{prob}@n$ shows higher bias in top-3 in comparison to top-10 and top-20 search results and the difference is even bigger between top-3 and top-20. For the NON-STEM field, the bias difference of only $\mathcal{R}ep_{prob}@3$ and $\mathcal{R}ep_{prob}@20$ is statistically significant with a lower difference in magnitude (effect size is 0.681) than the STEM field. Similarly, the bias differences of $\mathcal{E}xp_{prob}@3$ and $\mathcal{E}xp_{prob}@10$, $\mathcal{E}xp_{prob}@3$ and $\mathcal{E}xp_{prob}@20$ are statistically significant for the STEM field and the latter shows a higher difference (effect size of 0.857). Unlike the $\mathcal{R}ep_{prob}@n$, for the NON-STEM field, the bias difference of both $\mathcal{E}xp_{prob}@10$ and $\mathcal{E}xp_{prob}@20$, $\mathcal{E}xp_{prob}@3$ and $\mathcal{E}xp_{prob}@20$ are statistically significant with lower differences in magnitude (effect sizes are 0.291 and 0.621) than the STEM field.

In addition to these, the inaSpeechSegmenter model shows similar results with respect to the **RQ4**. For this model, in Table 7.5, the bias differences of $\mathcal{R}ep_{prob}@3$ and $\mathcal{R}ep_{prob}@10$ and $\mathcal{R}ep_{prob}@3$ and $\mathcal{R}ep_{prob}@20$ are statistically significant for the STEM field and again the latter shows a higher difference (effect size of 0.656), yet lower than Feed-Forward Gender Detector. This indicates that the STEM field shows higher bias in top-3 in comparison to top-10 and top-20 search results and the difference is even bigger between top-3 and top-20. For the NON-STEM field, the bias difference of only $\mathcal{R}ep_{prob}@3$ and $\mathcal{R}ep_{prob}@20$ is statistically significant with a slightly higher difference in magnitude (effect size is 0.658 instead of 0.656) than the STEM field, yet lower than the Feed-Forward Gender Detector. Similarly, the bias differences of $\mathcal{E}xp_{prob}@3$ and $\mathcal{E}xp_{prob}@10$, $\mathcal{E}xp_{prob}@3$ and $\mathcal{E}xp_{prob}@20$ are statistically significant for the STEM field and the latter shows a higher difference (effect size of 0.601), yet lower than the Feed-Forward Gender Detector (effect size of 0.857). Similar to the $\mathcal{R}ep_{prob}@n$, for the NON-STEM field, the bias difference of both $\mathcal{E}xp_{prob}@3$ and $\mathcal{E}xp_{prob}@10$, $\mathcal{E}xp_{prob}@3$ and $\mathcal{E}xp_{prob}@20$ are statistically significant with comparable differences in magnitude (effect sizes are 0.448 and 0.636) than the STEM field. Unlike the Feed-Forward Gender Detector, the bias difference in $\mathcal{E}xp_{prob}@10$ and $\mathcal{E}xp_{prob}@20$ is statistically not significant using the inaSpeechSegmenter model.

Regarding the **RQ5**, the source of bias is tracked to check whether it comes from the input data or the ranking algorithm – the two tailed *paired* t-tests applied on MB scores are statistically not significant regardless of the model, see Table 7.6 and Table 7.11. Unlike the MB scores, the two tailed *paired* t-tests applied on MAB scores show some statistically significant results both for the

Feed-Forward Gender Detector and inaSpeechSegmenter models. For the first model, in Table 7.6, the bias difference of only $\mathcal{R}ep_{prob}$@3 and $\mathcal{R}ep_{prob}$@200 is statistically significant for the STEM field with an effect size of 0.954 (noticeable difference in terms of magnitude). Unlike the STEM, for the NON-STEM field, all the bias differences using the MABs are statistically significant and the highest difference of bias in magnitude is between $\mathcal{R}ep_{prob}$@3 and $\mathcal{R}ep_{prob}$@200 with an effect size of 0.991 – higher than the STEM field as well. For the second model, in Table 7.11 again only the bias difference of $\mathcal{R}ep_{prob}$@3 and $\mathcal{R}ep_{prob}$@200 is statistically significant for the STEM field with an effect size of 0.522 that is lower than Feed-Forward Gender Detector. On the other hand, unlike the Feed-Forward Gender Detector model for the NON-STEM field, only the bias difference of $\mathcal{R}ep_{prob}$@3 and $\mathcal{R}ep_{prob}$@200 is statistically significant with an effect size of 0.830 that is lower. With respect to the **RQ5**, although both top results and the representative of the full corpus (which is the top-200 video search results) show bias, the magnitude of bias in the top results is higher than the full list. Thus, it can be inferred that the source of bias does not only come from the input data which is the indexed videos in the context of this chapter, but also from the ranking algorithm since in the top results there is a higher magnitude of *perceived* gender bias. In addition, the Feed-Forward Gender Detector model shows higher differences both for STEM and NON-STEM fields and for the NON-STEM field, not only the top-3 but also the results in top-10 and top-20 show high bias differences. Based on these findings, it seems that the ranking algorithm could also be blamed for the *perceived* gender bias results in online education using *YVRPs*.

Table 7.7 and Table 7.12 show the bias results for the STEM and NON-STEM majors. STEM majors indicate higher scores with the inaSpeechSegmenter model. The most biased STEM majors towards the male gender are Maths and Physics. The most biased NON-STEM majors towards the male gender is Politics and Psychology for the Feed-Forward Gender Detector model and only politics for the inaSpeechSegmenter. Sociology provides some negative scores for bothe models - biased towards the female. In Figure 7.2, the results show that both STEM and NON-STEM are overall biased towards the male (positive mean scores) but STEM is more biased both for the Feed-Forward Gender Detector and inaSpeechSegmenter models. Moreover, STEM shows higher bias using the inaSpeechSegmenter model and the difference between STEM and NON-STEM fields is also higher for it. This finding is also consistent with the aforementioned implication that the inaSpeechSegmenter model produces higher probability scores for the male gender. The error bar of both

STEM and NON-STEM fields are higher for the Feed-Forward Gender Detector model. Also, the error bar of the NON-STEM is higher than the STEM field. Figure 7.3 displays the bias scores of the measure, $\Delta_{\mathcal{R}ep_{prob}@n}$ for the STEM and NON-STEM fields using the Feed-Forward Gender Detector model, while Figure 7.5 using the inaSpeechSegmenter model. STEM field is more biased towards the male, i.e. more bubble points are in the upper-right quadrant, than the NON-STEM field using the inaSpeechSegmenter model. Same holds for the measure, $\Delta_{\mathcal{E}xp_{prob}@n}$, see Figure 7.4 and Figure 7.6. These observations are also consistent with the previous conclusions that the inaSpeechSegmenter model produces higher probability values for the male gender – favouring the male gender over female, please refer to the bias scores of STEM field in Table 7.4 and Table 7.9. For the NON-STEM field, the bubble points are only more dispersed with the inaSpeechSegmenter model. In Figure 7.5 and Figure 7.6, STEM is more biased than the NON-STEM towards the male – for the NON-STEM field there is no strong bias towards a specific gender group.

The results in Table 7.13 indicate that the STEM field displays a higher bias when the probability distributions scores are taken from the inaSpeechSegmenter. This is also consistent with the results in Figure 7.2. On the other hand, for the NON-STEM field there is no statistically significant bias difference based on the results in Table 7.14.

8. CONCLUSION & FUTURE WORK

Throughout this thesis, search engine bias through stance and ideological bias as well as online gender bias has been studied. In Chapter 3, new bias evaluation measures and a generalisable framework were introduced. Using these bias measures and the proposed framework, search results of Bing and Google were analysed with respect to stance and ideological bias. The results show that neither of the search engines are biased in terms of stances but both of them are ideologically biased – towards the liberal leaning. Then, in Chapter 4, search engine results were analysed with respect to bias and location in the scope of ideological bias. Chapter 5 aimed to track the source of stance and ideological bias by leveraging the state-of-the-art approaches to obtain labels automatically. In the second part, the thesis concentrates on gender bias in online education. In Chapter 6, two novel bias measures were proposed. Using these bias measures, YouTube video search results were analysed with respect to the unequal representation of male and female genders. It has been shown that the search results are biased towards the male and STEM field is more biased. Lastly, Chapter 7 aimed to investigate the source of gender bias in online education through YouTube videos. The findings indicate that the video search results are biased towards the male regardless of the annotation model and again STEM field is more biased.

As a future work, different approaches could be investigated to track the source of ideological bias in Bing and Google search results. Moreover, personalisation could be incorporated into the web search bias as well as *perceived* gender bias analysis.

BIBLIOGRAPHY

99Firms (2019). Search engine statistics. `https://99firms.com/blog/search-engine-statistics/#gref`. Accessed: 2019-09-06.

Aktolga, E. & Allan, J. (2013). Sentiment diversification with different biases. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, (pp. 593–602). ACM.

Alam, M. A. & Downey, D. (2014). Analyzing the content emphasis of web search engines. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, (pp. 1083–1086). ACM.

Alonso, O. & Mizzaro, S. (2012). Using crowdsourcing for trec relevance assessment. *Information processing & management*, *48*(6), 1053–1066.

Alonso, O., Rose, D. E., & Stewart, B. (2008). Crowdsourcing for relevance evaluation. In *SIGIR forum*, volume 42, (pp. 9–15).

Aslam, S. (2021). YouTube by the Numbers: Stats, Demographics & Fun Facts.

Baeza-Yates, R. (2016). Data and algorithmic bias in the web. In *Proceedings of the 8th ACM Conference on Web Science*, (pp. 1–1). ACM.

Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on facebook. *Science*, *348*(6239), 1130–1132.

Bargh, J. A., Gollwitzer, P. M., Lee-Chai, A., Barndollar, K., & Trötschel, R. (2001). The automated will: nonconscious activation and pursuit of behavioral goals. *Journal of personality and social psychology*, *81*(6), 1014.

Barman, C. R. (1997). Students' views of scientists and science: Results from a national study. *Science and Children*, *35*(1), 18.

Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Beutel, A., Chen, J., Doshi, T., Qian, H., Wei, L., Wu, Y., Heldt, L., Zhao, Z., Hong, L., Chi, E. H., et al. (2019). Fairness in recommendation ranking through pairwise comparisons. *arXiv preprint arXiv:1903.00780*.

Bodzin, A. & Gehringer, M. (2001). Breaking science stereotypes. *Science and Children*, *38*(4), 36.

Brooks, C., Gardner, J., & Chen, K. (2018). How gender cues in educational video impact participation and retention. International Society of the Learning Sciences, Inc.[ISLS].

Budak, C., Goel, S., & Rao, J. M. (2016). Fair and balanced? quantifying media bias through crowdsourced content analysis. *Public Opinion Quarterly*, *80*(S1), 250–271.

Ceci, L. (2021). Social network web visit share held by YouTube in the United Kingdom (UK) from January 2015 to October 2021.

Center, C. (2020). Tools and Guidance for Evaluating Bias in Instructional Materials.

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., & Tar, C. (2018). Universal sentence encoder. *arXiv preprint arXiv:1803.11175.*

Chelaru, S., Altingovde, I. S., & Siersdorfer, S. (2012). Analyzing the polarity of opinionated queries. In *European Conference on Information Retrieval*, (pp. 463–467). Springer.

Chelaru, S., Altingovde, I. S., Siersdorfer, S., & Nejdl, W. (2013). Analyzing, detecting, and exploiting sentiment in web queries. *ACM Transactions on the Web (TWEB)*, *8*(1), 6.

Chen, I.-X. & Yang, C.-Z. (2006). Position paper: A study of web search engine bias and its assessment. *IW3C2 WWW.*

Chen, L., Ma, R., Hannák, A., & Wilson, C. (2018). Investigating the impact of gender on rank in resume search engines. In *Proceedings of the 2018 chi conference on human factors in computing systems*, (pp. 1–14).

Chintalapati, N., Srinivas, V., & Daruri, K. (2016). Examining the use of YouTube as a Learning Resource in higher education: Scale development and validation of TAM model.

Chtouki, Y., Harroud, H., Khalidi, M., & Bennani, S. (2012). The impact of YouTube videos on the student's learning. In *2012 International Conference on Information Technology Based Higher Education and Training (ITHET)*, (pp. 1–4). IEEE.

Culpepper, J. S., Diaz, F., & Smucker, M. D. (2018). Research frontiers in information retrieval: Report from the third strategic workshop on information retrieval in lorne (swirl 2018). In *ACM SIGIR Forum*, volume 52, (pp. 46–47). ACM New York, NY, USA.

Delacre, M., Lakens, D., & Leys, C. (2017). Why psychologists should by default use welch's t-test instead of student's t-test. *International Review of Social Psychology*, *30*(1).

Demartini, G. & Siersdorfer, S. (2010). Dear search engine: what's your opinion about...?: sentiment analysis for semantic enrichment of web search results. In *Proceedings of the 3rd International Semantic Search Workshop*, (pp.4̃). ACM.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pretraining of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805.*

Diakopoulos, N., Trielli, D., Stark, J., & Mussenden, S. (2018). I vote for—how search informs our choice of candidate. *Digital Dominance: The Power of Google, Amazon, Facebook, and Apple, M. Moore and D. Tambini (Eds.)*, *22.*

Doukhan, D., Carrive, J., Vallet, F., Larcher, A., & Meignier, S. (2018). An open-source speaker gender detection framework for monitoring gender equality. In *Acoustics Speech and Signal Processing (ICASSP), 2018 IEEE International Conference on.* IEEE.

Draws, T., Tintarev, N., Gadiraju, U., Bozzon, A., & Timmermans, B. (2021). This Is Not What We Ordered: Exploring Why Biased Search Result Rankings Affect User Attitudes on Debated Topics. In *SIGIR 2021 - Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, (pp. 295–305). Association for Computing Machinery, Inc.

Dutton, W. H., Blank, G., & Groselj, D. (2013). *Cultures of the internet:*

the internet in Britain: Oxford Internet Survey 2013 Report. Oxford Internet Institute.

Dutton, W. H., Reisdorf, B., Dubois, E., & Blank, G. (2017). Search and politics: The uses and impacts of search in britain, france, germany, italy, poland, spain, and the united states.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness Through Awareness.

Elisa Shearer, K. E. M. (2018). News use across social media platforms 2018. https://www.journalism.org/2018/09/10/news-use-across-social-media-platforms-2018/.

Epstein, R. & Robertson, R. E. (2015). The search engine manipulation effect (seme) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Sciences*, *112*(33), E4512–E4521.

Epstein, R. & Robertson, R. E. (2017). A method for detecting bias in search rankings, with evidenceof systematic bias related to the 2016 presidential election. *Technical Report White Paper no. WP-17-02.*

Epstein, R., Robertson, R. E., Lazer, D., & Wilson, C. (2017). Suppressing the search engine manipulation effect (seme). *Proceedings of the ACM: Human-Computer Interaction*, *1*, 42.

EU (2017). Education.

Fang, Y., Si, L., Somasundaram, N., & Yu, Z. (2012). Mining contrastive opinions on political texts using cross-perspective topic model. In *Proceedings of the fifth ACM international conference on Web search and data mining*, (pp. 63–72). ACM.

Finin, T., Murnane, W., Karandikar, A., Keller, N., Martineau, J., & Dredze, M. (2010). Annotating named entities in twitter data with crowdsourcing. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, (pp. 80–88). Association for Computational Linguistics.

Gao, R. & Shah, C. (2019). How fair can we go: Detecting the boundaries of fairness optimization in information retrieval. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*, (pp. 229–236).

Gentzkow, M. & Shapiro, J. M. (2010). What drives media slant? evidence from us daily newspapers. *Econometrica*, *78*(1), 35–71.

Geyik, S. C., Ambler, S., & Kenthapadi, K. (2019). 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*, (pp. 2221–2231).

Ginger, G. & David, S. (2018). Google responds to trump, says no political motive in search results. https://www.reuters.com/article/us-usa-trump-tech-alphabet/google-responds-to-trump-says-no-political-motive-in-search-results-idUSKCN1LD1QP. Accessed: 2018-10-06.

Goldman, E. (2008). Search engine bias and the demise of search engine utopianism. In *Web Search* (pp. 121–133). Springer.

Gómez, E., Shui Zhang, C., Boratto, L., Salamó, M., & Marras, M. (2021). The Winner Takes it All: Geographic Imbalance and Provider (Un)fairness in Educational Recommender Systems. In *SIGIR 2021 - Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, (pp. 1808–1812). Association for Computing Machinery, Inc.

Good, C., Rattan, A., & Dweck, C. S. (2012). Why do women opt out? sense of belonging and women's representation in mathematics.

Grimes, D. R. (2016). Impartial journalism is laudable. but false balance is dangerous. `https://www.theguardian.com/science/blog/2016/nov/08/impartial-journalism-is-laudable-but-false-balance-is-dangerous`. Accessed: 2019-08-15.

Hannák, A., Wagner, C., Garcia, D., Mislove, A., Strohmaier, M., & Wilson, C. (2017). Bias in online freelance marketplaces: Evidence from taskrabbit and fiverr. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, (pp. 1914–1933).

Hardt, M., Price, E., Srebro, N., et al. (2016). Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, (pp. 3315–3323).

Hardt Google, M., Price, E., & Srebro, N. (2016). Equality of Opportunity in Supervised Learning. *Advances in neural information processing systems*, *29*, 3315–3323.

Hilton, J. L. & Von Hippel, W. (1996). Stereotypes Article in Annual Review of Psychology · .

Howard, J. & Ruder, S. (2018). Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Hu, D., Jiang, S., E. Robertson, R., & Wilson, C. (2019). Auditing the partisanship of google search snippets. In *The World Wide Web Conference*, (pp. 693–704).

Institute, A. P. (2014). The personal news cycle: How americans choose to get their news. *American Press Institute*.

InternetLiveStats2018 (2018). Internetlivestats. `http://www.internetlivestats.com/`. Accessed: 2018-10-06.

Järvelin, K. & Kekäläinen, J. (2000). Ir evaluation methods for retrieving highly relevant documents. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '00, (pp. 41–48)., New York, NY, USA. Association for Computing Machinery.

Joachims, T., Granka, L., Pan, B., Hembrooke, H., & Gay, G. (2005). Accurately Interpreting Clickthrough Data as Implicit Feedback.

Kallus, N. & Zhou, A. (2019). The fairness of risk scores beyond classification: Bipartite ranking and the xauc metric. *arXiv preprint arXiv:1902.05826*.

Kay, M., Matuszek, C., & Munson, S. A. (2015). Unequal representation and gender stereotypes in image search results for occupations. In *Conference on Human Factors in Computing Systems - Proceedings*, volume 2015-April, (pp. 3819–3828). Association for Computing Machinery.

Kizilcec, R. F. & Kambhampaty, A. (2020). Identifying course characteristics associated with sociodemographic variation in enrollments across 159 online courses from 20 institutions. *PloS one*, *15*(10), e0239766.

Kizilcec, R. F. & Saltarelli, A. J. (2019). Psychologically inclusive design:

Cues impact women's participation in stem education. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, (pp. 1–10).

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., et al. (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, *123*(1), 32–73.

Kulshrestha, J., Eslami, M., Messias, J., Zafar, M. B., Ghosh, S., Gummadi, K. P., & Karahalios, K. (2017). Quantifying search bias: Investigating sources of bias for political searches in social media. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, (pp. 417–432). ACM.

Kulshrestha, J., Eslami, M., Messias, J., Zafar, M. B., Ghosh, S., Gummadi, K. P., & Karahalios, K. (2018). Search bias quantification: investigating political bias in social media and web search. *Information Retrieval Journal*, 1–40.

Kwak, S. G. & Kim, J. H. (2017). Central limit theorem: the cornerstone of modern statistics. *Korean Journal of Anesthesiology*, *70*(2), 144.

Lahoti, P., Garimella, K., & Gionis, A. (2018). Joint non-negative matrix factorization for learning ideological leaning on twitter. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, (pp. 351–359).

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Lawson, N., Eustice, K., Perkowitz, M., & Yetisgen-Yildiz, M. (2010). Annotating large email datasets for named entity recognition with mechanical turk. In *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's Mechanical Turk*, (pp. 71–79). Association for Computational Linguistics.

Lee, C., Cho, K., & Kang, W. (2019). Mixout: Effective regularization to finetune large-scale pretrained language models. *arXiv preprint arXiv:1909.11299*.

Lipani, A., Piroi, F., & Yilmaz, E. (2021). Towards More Accountable Search Engines: Online Evaluation of Representation Bias.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

London, B., Rosenthal, L., Levy, S. R., & Lobel, M. (2011). *Basic and Applied Social Psychology*, *33*(4), 304–321.

McGuire, L., Mulvey, K. L., Goff, E., Irvin, M. J., Winterbottom, M., Fields, G. E., Hartstone-Rose, A., & Rutland, A. (2020). Stem gender stereotypes from early childhood through adolescence at informal science centers. *Journal of Applied Developmental Psychology*, *67*, 101109.

Mellebeek, B., Benavent, F., Grivolla, J., Codina, J., Costa-Jussa, M. R., & Banchs, R. (2010). Opinion mining of spanish customer comments with non-expert annotations on mechanical turk. In *Proceedings of the NAACL HLT 2010 workshop on Creating speech and language data with*

Amazon's mechanical turk, (pp. 114–121). Association for Computational Linguistics.

Mosbach, M., Andriushchenko, M., & Klakow, D. (2020). On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines. *arXiv preprint arXiv:2006.04884*.

Mowshowitz, A. & Kawaguchi, A. (2002a). Assessing bias in search engines. *Information Processing & Management*, *38*(1), 141–156.

Mowshowitz, A. & Kawaguchi, A. (2002b). Bias on the web. *Communications of the ACM*, *45*(9), 56–60.

Mullainathan, S. & Shleifer, A. (2005). The market for news. *American Economic Review*, *95*(4), 1031–1053.

Newman, N., Fletcher, R., Kalogeropoulos, A., Levy, D. A. L., & Nielsen, R. (2018). *Reuters institute digital news report 2018*, volume 2018. Reuters Institute for the Study of Journalism.

Newman, N., Fletcher, R., Kalogeropoulos, A., & Nielsen, R. (2019). *Reuters institute digital news report 2019*, volume 2019. Reuters Institute for the Study of Journalism.

Niwattanakul, S., Singthongchai, J., Naenudorn, E., & Wanapu, S. (2013). Using of jaccard coefficient for keywords similarity. In *Proceedings of the international multiconference of engineers and computer scientists*, volume 1, (pp. 380–384).

Noble, S. U. (2018). *Algorithms of Oppression: How search engines reinforce racism*. NYU Press.

Noonan, R. (2017). Women in STEM: 2017 Update.

Otterbacher, J., Bates, J., & Clough, P. (2017). Competent men and warm women: Gender stereotypes and backlash in image search results. In *Proceedings of the 2017 chi conference on human factors in computing systems*, (pp. 6620–6631).

Otterbacher, J., Checco, A., Demartini, G., & Clough, P. (2018). Investigating user perception of gender bias in image search: the role of sexism. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, (pp. 933–936).

Pan, B., Hembrooke, H., Joachims, T., Lorigo, L., Gay, G., & Granka, L. (2007). In google we trust: Users' decisions on rank, position, and relevance. *Journal of computer-mediated communication*, *12*(3), 801–823.

Piatek-Jimenez, K., Cribbs, J., & Gill, N. (2018). International journal of science education college students' perceptions of gender stereotypes: making connections to the underrepresentation of women in stem fields. *International Journal of Science Education*, *40*, 1432–1454.

ProCon.org (2018).

Räbiger, S., Gezici, G., Saygın, Y., & Spiliopoulou, M. (2018). Predicting worker disagreement for more effective crowd labeling. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, (pp. 179–188). IEEE.

Raji, I. D. & Buolamwini, J. (2019). Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, (pp. 429–435).

Reimers, N. & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Riegle-Crumb, C., Moore, C., & Buontempo, J. (2017). Shifting stem stereotypes? considering the role of peer and teacher gender. *Journal of Research on Adolescence*, *27*(3), 492–505.

Robertson, R. E., Jiang, S., Joseph, K., Friedland, L., Lazer, D., & Wilson, C. (2018). Auditing partisan audience bias within google search. *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW), 148.

Robertson, R. E., Lazer, D., & Wilson, C. (2018). Auditing the personalization and composition of politically-related search engine results pages. In *Proceedings of the 2018 World Wide Web Conference*, (pp. 955–965).

Saez-Trumper, D., Castillo, C., & Lalmas, M. (2013). Social media news communities: gatekeeping, coverage, and statement bias. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, (pp. 1679–1684). ACM.

Salmon, F. & Vallet, F. (2014). An effortless way to create large-scale datasets for famous speakers. In *LREC*, (pp. 348–352).

Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry*, *22*.

Sapiezynski, P., Zeng, W., E Robertson, R., Mislove, A., & Wilson, C. (2019). Quantifying the impact of user attentionon fair group representation in ranked lists. In *Companion Proceedings of The 2019 World Wide Web Conference*, (pp. 553–562).

Sarcona, C. (2019). Organic search click through rates: The numbers never lie. `https://www.zerolimitweb.com/organic-vs-ppc-2019-ctr-results-best-practices/`. Accessed: 2019-09-06.

Sedgwick, P. (2012). Multiple significance tests: the bonferroni correction. *Bmj*, *344*.

Singh, V. K., Chayko, M., Inamdar, R., & Floegel, D. (2020). Female librarians and male computer programmers? gender bias in occupational images on digital media platforms. *Journal of the Association for Information Science and Technology*, *71*(11), 1281–1294.

SmartSights2018 (2018). Search engine statistics 2018. `https://www.smartinsights.com/search-engine-marketing/search-engine-statistics/`. Accessed: 2018-10-06.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, *15*(1), 1929–1958.

Stokes, P. (2019). False media balance. `https://www.newphilosopher.com/articles/false-media-balance/`. Accessed: 2019-09-15.

Su, H., Deng, J., & Fei-Fei, L. (2012). Crowdsourcing annotations for visual object detection. In *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*.

Tavani, H. (2012). Search engines and ethics.

Vincent, N., Johnson, I., Sheehan, P., & Hecht, B. (2019). Measuring the importance of user-generated content to search engines. In *Proceedings* *of the International AAAI Conference on Web and Social Media*, volume 13, (pp. 505–516).

Vondrick, C., Patterson, D., & Ramanan, D. (2013). Efficiently scaling up crowdsourced video annotation. *International Journal of Computer Vision*, *101*(1), 184–204.

White, R. (2013). Beliefs and biases in web search. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, (pp. 3–12). ACM.

WISE (2018). 2018 workforce statistics - wise.

Yang, K. & Stoyanovich, J. (2017). Measuring fairness in ranked outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*, (pp.$\tilde{2}$2). ACM.

Yigit-Sert, S., Altingovde, I. S., & Ulusoy, Ö. (2016). Towards detecting media bias by utilizing user comments.

YouTube (2018). Youtube's impartiality practices - how does youtube work?

Yuen, M.-C., King, I., & Leung, K.-S. (2011). A survey of crowdsourcing systems. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, (pp. 766–773). IEEE.

Zehlike, M., Bonchi, F., Castillo, C., Hajian, S., Megahed, M., & Baeza-Yates, R. (2017). Fa* ir: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, (pp. 1569–1578). ACM.

Zhou, M. (2014). Gender difference in web search perceptions and behavior: Does it vary by task performance? *Computers & Education*, *78*, 174–184.