Noname manuscript No.
(will be inserted by the editor)

Evaluation Metrics for Measuring Bias in Search Engine Results

Gizem Gezici[1], Aldo Lipani[2], Yucel Saygin[1], and Emine Yilmaz[2]

[1]Sabanci University, Department of Computer Science and Engineering, Istanbul, Turkey
[2]University College London, Department of Computer Science, London, UK

Received: 4 February 2020 / Accepted: 7 December 2020

Abstract Search engines decide what we see for a given search query. Since many people are exposed to information through search engines, it is fair to expect that search engines are neutral. However, search engine results do not necessarily cover all the viewpoints of a search query topic, and they can be biased towards a specific view since search engine results are returned based on relevance, which is calculated using many features and sophisticated algorithms where search neutrality is not necessarily the focal point. Therefore, it is important to evaluate the search engine results with respect to bias. In this work we propose novel web search bias evaluation measures which take into account the rank and relevance. We also propose a framework to evaluate web search bias using the proposed measures and test our framework on two popular search engines based on 57 controversial query topics such as abortion, medical marijuana, and gay marriage. We measure the *stance bias* (in support or against), as well as the *ideological bias* (conservative or liberal). We observe that the stance does not necessarily correlate with the ideological leaning, e.g. a positive stance on abortion indicates a liberal leaning but a positive stance on Cuba embargo indicates a conservative leaning. Our experiments show that neither of the search engines suffers from stance bias. However, both search engines suffer from ideological bias, both favouring one ideological leaning to the other, which is more significant from the perspective of polarisation in our society.

Keywords Bias evaluation, Fair ranking, Search bias, Web Search

✉ Gizem Gezici
gizemgezici@sabanciuniv.edu

Address(es) of author(s) should be given

1 Introduction

Search engines have become an indispensable part of our lives. As reported by SmartSights (2018), 46.8% of the world population accessed the internet in 2017 and by 2021, the number is expected to reach 53.7%. According to InternetLiveStats (2018), currently on average 3.5 billion Google searches are done per day. These statistics indicate that search engines replaced traditional broadcast media and have become a *major* source of information "gatekeepers to the Web" for many people (Diaz, 2008). As information seekers search the Web more, they are also more influenced by Search Engine Result Pages (SERPs), pertaining to a wide range of areas (e.g., work, entertainment, religion, and politics). For instance, in the course of elections, it is known that people issue repeated queries on the Web about political candidates and events such as "democratic debate", "Donald Trump", "climate change" (Kulshrestha *et al.*, 2018). SERPs returned in response to these queries may influence the voting decisions as claimed by Epstein & Robertson (2015), who report that *manipulated* search rankings can change the voting preferences of undecided individuals at least by 20%.

Although search engines are widely used for seeking information, the majority of online users tend to believe that they provide *neutral* results, i.e. serving only as facilitators in accessing information on the Web (Goldman, 2008). However, there are counter examples to that belief as well. A recent dispute between the U.S. President Donald Trump and Google is such an example, where Mr. Trump accused Google of displaying only negative news about him when his name is searched to which Google responded by saying: "When users type queries into the Google Search bar, our goal is to make sure they receive the most relevant answers in a matter of seconds" and "Search is not used to set a political agenda and we don't bias our results toward any political ideology" (Ginger & David, 2018). In this work, we hope to shed some light on that debate, by not specifically concentrating on queries regarding Donald Trump but by conducting an in depth analysis of search answers to a broad set of controversial topics based on concrete evaluation measures.

Bias is defined with respect to balance in representativeness of Web documents retrieved from a database for a given query (Mowshowitz & Kawaguchi, 2002a). When a user issues a query to a search engine, documents from different sources are gathered, ranked, and displayed to the user. Assume that a user searches for *2016 presidential election* and the top-n ranked results are displayed. In such a search scenario, the retrieved results may favor some political perspectives over others and thereby fail to provide impartial knowledge for the given query as claimed by Mr. Trump, though without any scientific support. Hence, the potential *undue emphasis* of specific perspectives (or viewpoints) in the retrieved results lead to bias (Kulshrestha *et al.*, 2018). With respect to the definition of bias and the presented scenario, if there is an unbalanced representation, i.e. skewed or slanted distribution, of the viewpoints in a SERP, i.e. not only in political searches, towards the query's topic, then we consider this SERP as *biased* for the given search query.

Bias is especially important if the query topic is *controversial* having opposing views, in which case it becomes more critical that search engines are supposed to return results with a *balanced* representation of different perspectives which implies that they do not favour one specific perspective over another. Otherwise, this may dramatically affect public as in the case of elections leading to polarisation in society for *controversial* issues. On the other hand, returning an unbalanced representation of distinct viewpoints is not sufficient to claim that the search engine's ranking algorithm is biased. One reason for a skewed SERP could be due to the corpus itself, i.e. if documents indexed and returned for a given topic come from a slanted distribution, meaning that the ranking algorithm returns a biased result set due to a biased corpus. To differentiate the algorithmic vs corpus bias, one needs to investigate the source of bias in addition to the skewed list analysis of the top-n search results. However, the existence of bias, regardless of being corpus or algorithmic bias, would still conflict with the expectation that an IR system should be fair, accountable, and transparent (Culpepper *et al.*, 2018). Furthermore, it was reported that people are more susceptible to bias when they are unaware of it (Bargh *et al.*, 2001), and Epstein *et al.* (2017) showed that alerting users about bias can be effective in suppressing search engine manipulation effect (SEME). Thus, search engines should at least inform their users about the bias and decrease the possible SEME by making themselves more accountable, thereby alleviating the negative effects of bias and serving only as facilitators as they generally claim to be. In this work, we aim to serve that purpose by proposing a search bias evaluation framework taking into account the rank and relevance [1] of the SERPs. Our contributions in this work can be summarised as follows:

1. We propose a *new generalisable search bias evaluation framework* to measure bias in SERPs by quantifying two different types of bias on content which are stance bias and ideological bias.
2. We present *three novel fairness-aware measures of bias* that do not suffer from the limitations of the previously presented bias measures, 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) which are explained in Section 3.2 in detail.
3. We apply the proposed framework to *measure the stance and ideological bias* not only in political searches but searches related to a wide range of controversial topics; including but not limited to education, health, entertainment, religion and politics on Google and Bing *news* search results.
4. We also utilise our framework to *compare the relative bias* for queries from various controversial issues on two popular search engines: Google and Bing news search.

We would like to note that we distinguish the stance and ideological leaning in SERPs. The stance in a SERP for a query topic could be in favor or against

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

the topic, whereas the ideological leaning in a SERP stands for the specific ideological group as conservatives or liberals that supports the corresponding topic. Hence, the stance in a SERP does not directly imply the ideological leaning. For example, given two controversial queries, "abortion" and "Cuba embargo", a SERP could have a positive stance for the topic of abortion, indicating a liberal leaning, while a positive stance for the topic of Cuba embargo indicates a conservative leaning. Therefore looking at the stance of the SERPs for controversial issues is not enough and could even be misleading in determining the ideological bias. We demonstrate how the proposed framework can be used to quantify bias in the SERPs of search engines (in this case Bing and Google) in response to queries related to *controversial* topics. Our analysis is mainly two-fold where we first evaluate stance bias in SERPs, and then use this evaluation as a proxy to quantify ideological bias asserted in the SERPs of the search engines.

In this work, via the proposed framework, we aim to answer the following research questions:

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

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

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?

We address these research questions for 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. In order to answer RQ1, we measure the degree of deviation of the ranked SERPs from an *ideal* distribution, where different stances are *equally* likely to appear. To detect bias which results from the unbalanced representation of distinct perspectives, we label the documents' stances with crowd-sourcing and use these labels for stance bias evaluation. In this paper we focus on a particular kind of bias, *statistical parity* or more generally known as *equality of outcome*, i.e. given a population divided into groups, the groups in the output of the system should be equally represented. This is in contrast with the other popular measure generally known as *equality of opportunity*, i.e. given a population divided into groups, the groups in the output should be represented based on their proportion in the population namely, base rates. For choosing the *equality of outcome*, we have mainly two reasons. First, in the context of the controversial topics, not all of the corresponding debate questions (queries) have certain answers based on scientific facts. Second, the identification of the stance for the full ranking list, i.e. which is a fair representative set of the indexed documents, is too expensive to get annotated through crowd-sourcing. Thus, this choice of *ideal* ranking makes the experiments feasible. To address RQ2, we compare the stance bias in the SERPs of the two search engines to see if they show similar level of bias for the corresponding controversial topics. RQ3 is naturally answered by assigning an ideological leaning label to each query topic as conservative or liberal depending on which ideology favors the proposition in the query. We further interpret the document stance labels in conservative-to-liberal ideology [2] space and transform these stance labels into ideological leanings according to the assigned leaning labels of the corresponding topics. We note that conservative-to-liberal ideology space does not only stand for political parties. In this context, we accept these ideology labels as having a more conservative/liberal viewpoint towards a given controversial topic as similarly fulfilled by Lahoti *et al.* (2018) for three popular controversial topics of *gun control*, *abortion*, and *obamacare* in Twitter domain.

For instance, the topic of *abortion* has the query of *Should Abortion Be Legal?* Since mostly liberals support the proposition in this query, liberal leaning is assigned to abortion. The stance labels of the retrieved documents towards the query are transformed into ideological leanings as follows. If a document has the pro stance which means that it supports the asserted proposition, then its ideological leaning is liberal; if it has the against stance, its leaning is conservative.

In our bias evaluation framework, we concentrate on the top-10 SERPs coming from the *news* sources to investigate two major search engines (Bing and Google) in terms of bias. We deliberately use *news* SERPs for our experiments since they often exhibit a specific view towards a topic (Alam & Downey, 2014). Recent studies (Sarcona, 2019; 99Firms, 2019) show that on average more than 70% of all the clicks are in the first page results, thus we only focus on the top-10 results to show the existence of bias. Experiments show that there is no statistically significant difference of *stance bias* in magnitude measured across the two search engines, meaning that they do not favour one specific stance over other. However, we should stress that stance bias results need to be taken with a grain of salt as demonstrated through the abortion and Cuba embargo query examples. Polarisation of the society is mostly on ideological leanings, and our second phase of experiments show that there is statistically significant difference of *ideological bias*, where both search engines favour one ideological leaning over other.

The remainder of the paper is structured as follows. In Section 2 we give the related work and the search bias evaluation framework is proposed in Section 3. In Section 4 we detail the experimental setup, and present the results. Then, we discuss the results in Section 5. In Section 6 we present the limitations of this work, and we conclude in Section 7.

2 Background & Related Work

In recent years, bias analysis in SERPs of search engines has attracted a lot of interest (Baeza-Yates, 2016; Mowshowitz & Kawaguchi, 2002b; Noble, 2018; Pan *et al.*, 2007; Tavani, 2012) due to the concerns that search engines may

[2] We are referring to the notion of ideology perceived by the crowd workers.

manipulate the search results influencing users. The main reason behind these concerns is that search engines have become the fundamental source of information (Dutton *et al.*, 2013), and surveys from Pew (2014) and Reuters (2018) found that more people obtain their news from search engines than social media. The users reported higher trust on search engines for the accuracy of information (Newman *et al.*, 2018, 2019; Elisa Shearer, 2018) and many internet-using US adults even use search engines to fact-check information (Dutton *et al.*, 2017).

To figure out how this growing usage of search engines and trust in them might have undesirable effects on public, and what could be the methods to measure those effects, in the following we review the research areas related first to automatic stance detection, then to fair ranking evaluation, and lastly to search bias quantification.

2.1 Opinion Mining and Sentiment Analysis

A form of Opinion Mining related to our work is Contrastive Opinion Modeling (COM). Proposed by Fang *et al.* (2012), in COM, given a political text collection, the task is to present the opinions of the distinct perspectives on a given query topic and to quantify their differences with an unsupervised topic model. COM is applied on debate records and headline news. Differently from keyword analysis to differentiate opinions using topic modelling, we compute different IR metrics from the content of the news articles to evaluate and compare the bias in the SERPs of two search engines. Aktolga & Allan (2013) consider the sentiment towards controversial topics and propose different diversification methods based on the topic sentiment. Their main aim is to diversify the retrieved results of a search engine according to various sentiment biases in blog posts rather than measure bias in the SERPs of *news* search engines as we do in this work.

Demartini & Siersdorfer (2010) exploit automatic and lexicon-based text classification approaches, Support Vector Machines and SentiWordNet respectively to extract sentiment value from the textual content of SERPs in response to controversial topics. Unlike us, Demartini & Siersdorfer (2010) only use this sentiment information to compare opinions in the retrieved results of three commercial search engines without measuring bias. In this paper, we propose a new bias evaluation framework with robust bias measures to systematically measure bias in SERPs. Chelaru *et al.* (2012) focus on queries rather than SERPs and investigate if the opinionated queries are issued to search engines by computing the sentiment of suggested queries for controversial topics. In a follow-up work (Chelaru *et al.*, 2013), authors use different classifiers to detect the sentiment expressed in queries and extend the previous experiments with two different use cases. Instead of queries, our work analyses the SERPs in *news* domain, therefore we need to identify the stance of the news articles. Automatically obtaining article stances is beyond the scope of this work, thus we use crowd-sourcing.

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) that are related to Normalized Discounted Cumulative Gain (NDCG) through the use of logarithmic discounting for regularization which is inspired from NDCG as also stated in the original paper. Researchers use these metrics to check if there exists a systematic discrimination against a group of individuals, when there are only two different groups as a protected (g_1) and an unprotected group (g_2) in a ranking. In other words, researchers quantify the relative representation of g_1 (the protected group), whose members share a characteristic such as race or gender that cannot be used for discrimination, in a ranked output. The definitions of these three proposed measures can be rewritten as follows:

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

where $f(r)$ is a general definition of an evaluation measure for a given ranked list of documents, i.e. a SERP, whereas f_{g_1} is specifically for the protected group of g_1. In this definition, Z is a normalisation constant, r is the ranked list of the retrieved SERP and $|r|$ is the size of this ranked list, i.e. number of documents in the ranked list. Note that, i is deliberately incremented by 10, to compute *set-based fairness* at discrete values as top-10, top-20 etc., instead of 1 as usually done in IR for the proposed measures to show the correct behaviour with bigger sample sizes. The purpose of computing the *set-based fairness* to express that being fair at higher positions of the ranked list is more important, e.g. top-10 vs. top-100.

In the rewritten formula, d_{g_1} defines a distance function between the expected probability to retrieve a document belonging to g_1, i.e. in the overall population, and its observed probability at rank i to measure the systematic bias. These probabilities turn out to be equal to P@n:

$$\mathrm{P}_{g_1}@n = \frac{1}{n} \sum_{i=1}^{n} [j(r_i) = g_1], \tag{2}$$

when computed over g_1 at cut-off value $|r|$ and i for the three proposed measures as below. In this formula, n is the number of documents considered in r as a cut-off value, and r_i is defined as the document in r retrieved at rank i. Note that, $j(r_i)$ returns the label associated to the document r_i specifying its group as g_1 or g_2. Based on this, $[j(r_i) = g_1]$ refers to a conditional statement which returns 1 if the document r_i is the 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:

1. rND measure focuses on the protected group (g_1). If we were to compute f at steps of 1 with the given equal desired proportion of the two groups as 50:50, then the distance function of rND, denoted as d_{g_1} would always give a value of 0.5 for the first retrieved document, where $i = 1$. This will always be the case, no matter which group this document belongs to, e.g. *pro* or *against* in our case. This is caused by d_{g_1} of rND through the use of its absolute value in Eq. (1). In our case, this holds when $i = 1, 2, 4$ and $r = 10$ where we measure bias in the top-10 results. This is in fact avoided in the original paper (Yang & Stoyanovich, 2017) by computing f at steps of 10 as top-10, top-20 etc. rather than the steps of 1 as it is usually done in IR which gives more meaningful results in our evaluation framework.
2. rKL measure cannot differentiate between biases of equal magnitude, but in opposite directions with the given equal desired proportion of the two groups as 50:50, i.e. it cannot differentiate bias towards *conservative*, or *liberal* in our case. Also, in IR settings it is not as easy to interpret the computed values from the KL-divergence (denoted as d_{g_1} for rKL) compared to our measures since our measures are based on the standard utility-based IR measures. Furthermore, KL-divergence tends to generate larger distances for small datasets, thus it could compute larger bias values in the case of only 10 documents, and this situation may become even more problematic if we measure bias for less number of documents, e.g. top-3, top-5 for a more fine-grained analysis. In the original paper, this disadvantage is alleviated by computing the rKL values also at discrete points of steps 10 instead of 1.
3. rRD measure does not treat the protected and unprotected groups (g_1 and g_2) symmetrically as stated in the original paper, which is not applicable to our framework. Our proposed measures treat g_1 and g_2 equal since we have two protected groups; *pro* and *against* for stance bias, *conservative* and *liberal* for ideological bias to measure bias in search settings. Moreover, rRD is only applicable in special conditions when g_1 is the minority group

in the underlying population as also declared by the authors, while we do not have such constraints for our measures in the scope of search bias evaluation.

4. These measures focus on differences in the relative representation of g_1 between distributions. Therefore, from a general point of view, most probably more samples are necessary for these measures to show the expected behavior and work properly. In the original paper, experiments are fulfilled with three different datasets, one is synthetic which includes 1000 samples and two are real datasets which include 1000 and 7000 samples to evaluate bias with these measures, while we have only 10 samples for query-wise evaluation. This is probably because these measures were mainly devised for the purpose of measuring bias in ranked outputs instead of search engine results; none of these datasets contain search results either.
5. These measures are difficult to use in practice, since they rely on a normalization term, Z that is computed stochastically, i.e. as the highest possible value of the corresponding bias measure for the given number of documents n and protected group size $|g_1|$. In this paper, we rely on standard statistical tests, since they are easier to interpret, provide confidence intervals, and have been successfully used to investigate inequalities in search systems previously by Chen *et al.* (2018).
6. These measures do not consider relevance which is a fundamental aspect when evaluating bias in search engines. For example, as in our case, when searching for a controversial topic, if the first retrieved document is about a news belonging to g_1 but its content is not relevant to the searched topic, then these measures would still consider this document as positive for g_1. However, this document has absolutely no effect on providing an unbiased representation of the controversial topic to the user. This is because these metrics were devised particularly for evaluating bias in the ranked outputs instead of SERPs.

Although the proposed measures by Yang & Stoyanovich (2017) are valuable in the context of measuring bias in ranked outputs where the individuals are being ranked and some of these individuals are the members of the protected group (g_1), these measures have the aforementioned limitations. These limitations are particularly visible for content bias evaluation where the web documents are being ranked by search engines in a typical IR setting. In this paper we address these limitations by proposing a family of fairness-aware measures with the main purpose of evaluating content bias in SERPs, based on standard utility-based IR evaluation measures.

Zehlike *et al.* (2017), based on Yang & Stoyanovich (2017)'s work, propose an algorithm to test the statistical significance of a fair ranking. Beutel *et al.* (2019) propose a pairwise fairness measure for recommender systems. However, the authors, unlike us, measure fairness on personalized recommendations and do not consider relevance, while we work in an unpersonalized information retrieval setting and we do consider relevance. Kallus & Zhou (2019) investigate the fairness of predictive risk scores as a bipartite ranking task, where the main goal is to rank positively labelled examples above negative ones. However, their measures of bias based on the area under the ROC curve (AUC) are agnostic from the rank position at which a document has been retrieved.

2.3 Quantifying Search Engine Biases

Although the search engine algorithms are not transparent and available to external researchers, algorithm auditing techniques provide an effective means for systematically evaluating the results in a controlled environment (Sandvig *et al.*, 2014). Prior works leverage LDA-variant unsupervised methods and crowd-sourcing to analyse bias in content, or URL analysis for indexical bias.

Saez-Trumper *et al.* (2013) propose unsupervised methods to characterise different types of biases in online news media and in their social media communities by also analysing political perspectives of the news sources. Yigit-Sert *et al.* (2016) investigate media bias by analysing the user comments along with the content of the online news articles to identify the latent aspects of two highly polarising topics in the Turkish political arena. Kulshrestha *et al.* (2017) quantify bias in social media by measuring the bias of the author of a tweet, while in Kulshrestha *et al.* (2018), bias in web search is quantified through a URL analysis for Google in political domain without any SERP content analysis. In our work, we consider the Google and Bing SERPs from news sources such as NY-Times, and BBC news in order to quantify bias through content analysis.

In addition to the unsupervised approaches, crowd-sourcing is a widely used mechanism to analyse bias in content. Crowd-sourcing is a common approach for labelling tasks in different research areas such as image & video annotation (Krishna *et al.*, 2017; Vondrick *et al.*, 2013), object detection (Su *et al.*, 2012), named entity recognition (Lawson *et al.*, 2010; Finin *et al.*, 2010), sentiment analysis (Räbiger *et al.*, 2018) and relevance evaluation (Alonso *et al.*, 2008; Alonso & Mizzaro, 2012). Yuen *et al.* (2011) provide a detailed survey of crowd-sourcing applications. As Yuen *et al.* (2011) suggest, crowd-sourcing can also be used for gathering opinions from the crowd. Mellebeek *et al.* (2010) use crowd-sourcing to classify Spanish consumer comments and show that non-expert Amazon Mechanical Turk (MTurk) annotations are viable and cost-effective alternative to expert ones. In this work, we use crowd-sourcing for collecting opinions of the public not about consumer products but controversial topics.

Apart from the content bias, there is another research area, namely indexical bias. Indexical bias refers to the bias which is displayed in the selection of items, rather than in the content of retrieved documents, namely content bias (Mowshowitz & Kawaguchi, 2002b). Mowshowitz & Kawaguchi (2002a, 2005) quantify instead only indexical bias by using precision and recall measures. Moreover, the researchers approximate the *ideal* (i.e. norm) by the distribution produced by a collection of search engines to measure bias. Yet, this may not be a *fair* bias evaluation procedure since the *ideal* itself should be *unbiased*, whereas the SERPs of search engines may actually contain *bias*. Similarly, Chen & Yang (2006) use the same method in order to quantify indexical and content bias, however, content analysis was performed by representing the SERPs with a weighted vector with different HTML tags without an in-depth analysis of the textual content. In this work, we evaluate content bias by analysing the textual contents of the Google and Bing SERPs, and we do not generate the *ideal* relying on the SERPs of other search engines in order to measure bias in a more *fair* way. In addition to the categorisation of the content and indexical bias analysis, prior methods used in auditing algorithms to quantify bias can also be divided into three main categories as *audience-based*, *content-based*, and *rater-based*. *Audience-based* measures focus on identifying the political perspectives of media outlets and web pages by utilising the interests, ideologies, or political affiliations of its users, e.g., likes and shares on Facebook (Bakshy *et al.*, 2015), based on the premise that readers follow the news sources that are closest to their ideological point of view (Mullainathan & Shleifer, 2005). Lahoti *et al.* (2018) model the problem of ideological leaning of social media users and media sources in the liberal-conservative ideology space on Twitter as a constrained non-negative matrix-factorisation problem. *Content-based* measures exploit linguistic features in textual content; Gentzkow & Shapiro (2010) extract frequent phrases of the different political partisans (Democrats, Republicans) from the Congress Reports. Then, the researchers come with the metric of media slant index to measure US newspapers' political leaning. Finally, rater-based methods also exploit textual content and can be evaluated under the content-based methods. Unlike the content-based, the *rater-based* methods use ratings of people for the sentiment, partisan or ideological leaning of content instead of analysing the textual content linguistically. Rater-based methods generally leverage crowd-sourcing to collect the labels for the content analysis. For instance, Budak *et al.* (2016) quantify bias (partisanship) in US news outlets (newspapers and 2 political blogs) for 15 selected queries related to a wide range of controversial issues about which Democrats and Republicans argue. The researchers use MTurk as a crowd-sourcing platform to obtain the topic and political slant labels, i.e. being positive towards Democrats or Republicans, of the articles. Similarly, Epstein & Robertson (2017) use crowd-sourcing to score individual search results and Diakopoulos *et al.* (2018) make use of the MTurk platform, i.e. rater-based approach, to get labels for the Google SERP websites by focusing on the content and apply an audience-based approach through utilising the prior work of Bakshy *et al.* (2015) specifically for quantifying partisan bias. Our work follows a rater-based approach by making use of the MTurk platform for crowd-sourcing to analyse web search bias through stances and ideological leanings of the news articles instead of partisan bias in the textual contents of the SERPs.

There have been endeavors to audit partisan bias on web search. Diakopoulos *et al.* (2018) present four case studies on Google search results and to quantify partisan bias in the first page, they collect SERPs by issuing complete candidate names of the 2016 US presidential election as queries and utilise crowd-sourcing to obtain the sentiment scores of the SERPs. They found that Google presented a higher proportion of negative articles for Republican candidates than the Democratic ones. Similarly, Epstein & Robertson (2017) present a case study for the election and use a browser extension to collect Google and Yahoo search data for the election-related queries, then use crowd-sourcing to score the SERPs. The researchers also found a left-leaning bias and Google was more biased than Yahoo. In their follow-up work, they found a small but significant ranking bias in the standard SERPs but not due to personalisation (Robertson *et al.*, 2018a). Similarly, researchers audit Google search after Donald Trump's Presidential inauguration with a dynamic set of political queries using auto-complete suggestions (Robertson *et al.*, 2018b). Hu *et al.* (2019) conduct an algorithm audit and construct a specific lexicon of partisan cues for measuring political partisanship of Google Search snippets relative to the corresponding web pages. They define the corresponding difference as bias for this particular use case without making a robust search bias evaluation of SERPs from the user's perspective. In this work, we introduce novel fairness-aware IR measures which involve rank information to evaluate content bias. For this, we use crowd-sourcing to obtain labels of the *news* SERPs returned towards the queries related to a wide-range of controversial topics instead of only political ones. With our robust bias evaluation measures, our main aim is to audit ideological bias in web search rather than solely partisan bias.

Apart from partisan bias, recent studies have investigated different types of bias for various purposes. Chen *et al.* (2018) investigate gender bias in the various resume search engines, which are platforms that help recruiters to search for suitable candidates and use statistical tests to examine two types of indirect discrimination: individual and group fairness. Similarly in another research study, authors investigate gender stereotypes by analyzing the gender distribution in image search results retrieved by Bing in four different regions (Otterbacher *et al.*, 2017). Researchers use the query of 'person' and the queries related to 68 character traits such as 'intelligent person', and the results show that photos of women are more often retrieved for 'emotional' and similar traits, whereas 'rational' and related traits are represented by photos of men. In a follow-up work, researchers conduct a controlled experiment via crowd-sourcing with participants from three different countries to detect bias in image search results (Otterbacher *et al.*, 2018). Demographic information along with measures of sexism are analysed together and the results confirm that sexist people are less likely to detect and report gender biases in the search results.

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 metrics proposed by Yang & Stoyanovich (2017), of which we specified the limitations in Section 2.2, and evaluated their procedure using simulations with application to LinkedIn Talent Search. Vincent *et al.* (2019) measure the dependency of search engines on user-created content to respond to queries using Google search and Wikipedia articles. In another work, researchers propose a novel metric that involves users and their attention for auditing group fairness in ranked lists (Sapiezynski *et al.*, 2019). Gao & Shah (2019) propose a framework that effectively and efficiently estimate the solution space where fairness in IR is modelled as an optimisation problem with fairness constraint. Same researchers work on top-k diversity fairness ranking in terms of statistical parity and disparate impact fairness and propose entropy-based metrics to measure the topical diversity bias presented in SERPs of Google using clustering instead of a labelled dataset with group information (Gao & Shah, 2020). Unlike to their approach, our goal is to quantify search bias in SERPs rather than topical diversity. For this, we use a crowd-labelled dataset, thereby to evaluate bias from the user's perspective with stance and ideological leanings of the documents.

In this context, we focus on proposing a new search bias evaluation procedure in ranked lists to quantify bias in the *news* SERPs. With the proposed robust fairness-aware IR measures, we also compare the relative bias of the two search engines through incorporating relevance and ranking information into the procedure without tracking the source of bias as discussed in Section 1. Our procedure can be used for the source of bias analysis as well which we leave as future work.

3 Search Engine Bias Evaluation Framework

In this section we describe our search bias evaluation framework. Then, we present the measures of bias and the proposed protocol to identify search bias.

3.1 Preliminaries

Our first aim is to detect bias with respect to the distribution of stances expressed in the contents of the SERPs.

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 issued 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.

For our analyses, we deliberately use recent *controversial* topics in US that are the real debatable ones rather than the topics being possibly exposed to false media balance, which occurs when the media present opposing viewpoints as being more equal than the evidence supports, e.g. Flat Earth debate (Grimes, 2016; Stokes, 2019). Our topic set contains abortion, illegal immigration, gay marriage, and similar *controversial* topics which comprise opposing points of view since complicated concepts concerning the identity, religion, political or ideological leaning are the actual points where search engines are more likely to provide biased results (Noble, 2018) and influence people dramatically.

Our second aim is to detect bias with respect to the distribution of ideological leanings expressed in the contents of the SERPs. We do this by associating each query $q \in \mathcal{Q}$ belonging to a controversial topic to one *current* ideological leaning. Then, combining the stances for each r_i and the associated ideological leaning of q we can measure the ideological bias of the content of a given SERP, e.g. if a topic belongs to a specific ideology and a document retrieved for this topic has a pro stance, we consider this document to be biased towards this ideology. We define the ideological leaning of q as $j(q)$. An ideological leaning can have the following values: *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.

For reference, Table 1 shows a summary of all the symbols, functions and labels used in this paper.

3.2 Measures of Bias

Based on the aforementioned definition provided in Section 1, bias can be quantified by measuring the degree of deviation of the distribution of documents from the *ideal* one. To give a broad definition of an ideal list poses problems; but in the scope of this work for *controversial* topics, we can mention the existence of bias in a ranked list retrieved by a search engine if the presented information *significantly deviates* from true likelihoods (White, 2013). As justified in Section 1, in the scope of this work we focus on *equality of output*, thus we accept the true likelihoods of different views as *equal* rather than computing them from the corresponding base rates. Therefore using the proposed definition reversely, we can assume that the *ideal* list is the one that minimises the difference between two opposing views, which we indicate here

Table 1 Symbols, functions, and labels used throughout the paper

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.
$\lvert r \rvert$	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.

as 👍 and 👎 in the context of stances. Formally, we measure the *stance* bias in a SERP r as follows:

$$\beta_f(r) = f_{👍}(r) - f_{👎}(r), \tag{3}$$

where f is a function that measures the likelihood of r in satisfying the information need of the user about the view 👍 and the view 👎. We note that *ideological* bias is measured in the same way by transforming the stances of the documents into ideological leanings which will be explained in Section 4.2. Before defining f, from Eq. (3), we define the mean bias (MB) of a search engine s as:

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

An unbiased search engine would produce a mean bias of 0. A limitation of MB is that if a search engine is biased towards the 👍 view on one topic and bias towards the 👎 view on another topic, these two contributions will cancel each other out. In order to avoid this limitation we also define the mean absolute bias (MAB), which consists in taking the absolute value of the bias for each r. Formally, this is defined as follows:

$$\mathrm{MAB}_f(s, \mathcal{Q}) = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} |\beta_f(s(q))|. \tag{4}$$

An unbiased search engine produces a mean absolute bias of 0. Although this measure defined in Eq. (4) solves the limitation of MB, MAB says nothing about towards which view the search engine is biased, making these two measures of bias complementary.

In IR the likelihood of r in satisfying the information need of users is measured via retrieval evaluation measures. Among these measures we selected 3 *utility-based* evaluation measures. This class of evaluation measures quantify r in terms of its worth to the user and are normally computed as a sum of the information gain summed over the relevant documents retrieved by r. The 3 IR evaluation measures used in the following experiments are: P@n, RBP, and DCG@n.

P@n for the 👍 view is formalised as in Eq. (2). However, differently from the previous definition of $j(r_i)$ where the only possible outcomes are g_1 and g_2 for the document r_i, here j can return any of the label associated to a stance (👍, ✊, 👎, and ✖). Hence, only pro and against documents, that are relevant to the topic, are taken into account, since $j(r_i)$ returns *neutral* and *not-relevant* when otherwise. Substituting Eq. (2) to Eq. (3) we obtain the first measure of bias:

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

The main limitation of this measure of bias is that it has a weak concept of ranking, i.e. the first n documents contribute equally to the bias score. The next two evaluation measures overcome this issue by defining discount functions.

RBP weights every document based on the coefficients of a normalised geometric series with value $p \in]0,1[$, where p is a parameter of RBP. Similarly to what is done for P@n, we reformulate RBP to measure bias as follows:

$$\mathrm{RBP}_{\text{👍}} = (1-p) \sum_{i=1} p^{i-1} [j(r_i) = \text{👍}]. \tag{5}$$

Substituting Eq. (5) to Eq. (3) we obtain:

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

DCG@n, instead, weights each document based on a logarithmic discount function. Similarly to what is done for P@n and RBP, we reformulate DCG@n to measure bias as follows:

$$\mathrm{DCG}_{\text{👍}}@n = \sum_{i=1}^{n} \frac{1}{\log(i+1)} [j(r_i) = \text{👍}]. \tag{6}$$

Substituting Eq. (6) to Eq. (3) we obtain:

$$\beta_{\mathrm{DCG@}n}(r) = \sum_{i=1}^{n} \frac{1}{\log(i+1)} \left([j(r_i) = \text{👍}] - [j(r_i) = \text{👎}] \right) \tag{7}$$

Since we are evaluating web-users, for P@n and DCG@n we set $n = 10$ and for RBP we set $p = 0.8$. This last formulation (Eq. (7)), although it looks similar to the rND measure, it does not suffer from the four limitations introduced in Section 2.2. In particular all these presented measures of bias: 1) do not focus on one group; 2) use a binary score associated to the document stance or ideological leaning, similar to the way these measures are used in IR when considering relevance; also like in IR 3) can be computed at each rank; 4) exclude non-relevant documents from the measurement of bias and; the framework 5) provides various user models associated to the 3 IR evaluation measures: P@n, DCG@n, and RBP.

3.3 Quantifying Bias

Using the measures of bias defined in the previous section we quantify the bias of the two search engines, Bing and Google using the *news versions* of these search engines. Then, we compare them thereof. Following, we describe each step of the proposed procedure used to quantify bias in SERPs.

- **News Articles in SERPs.** We obtained the controversial queries issued for searching from ProCong.org [2018] and applied some filtering steps on the initial query set. After filtering, the final query set size became 57. We submitted each query in the final query set to the US News search engines of Google and Bing using a US proxy. Then, we extracted the whole corpus returned by both engines in response to all the queries in the set. Note that the data collection process was done in a controlled environment such that the queries are sent to the search engines at the same time. For more details about the selection of the queries and crawling the SERPs, please refer to the previous phase of our analysis. After having crawled all the SERPs returned from both engines and extracted their contents, we annotated the top 10 documents. We obtained the stance label of each document with respect to the queries via crowd-sourcing. To label the ideological leaning of queries, we also used crowd-sourcing. To obtain the ideologies of documents, we transformed the stance labels into ideologies based on the ideological leaning of their corresponding queries. The details about our crowdsourcing campaigns as well as the transformation process can also be found in the first phase of our analysis.
- **Bias Evaluation.** We compute the bias measures for every SERP with all three IR-based measures of bias: P@n, RBP, and DCG@n. We then aggregate the results using the two measures of bias, MB and MAB.
- **Statistical Analysis.** To identify whether the bias measured is not a byproduct of randomness, we compute a one-sample t-test: the null hypothesis is that no difference exists and that the true mean is equal to zero. If this hypothesis is rejected, hence there is a significant difference and we claim that the evaluated search engine is biased. Then, we compare the difference in bias measured across the two search engines using a two-tailed paired t-test: the null hypothesis is that the difference between the two true means is equal to zero. If this hypothesis is rejected, hence

Table 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?

there is a significant difference, we claim that there is a difference in bias between the two search engines.

4 Experimental Setup

In this section we provide a description of our experimental setup based on the proposed method as defined in Section 3.3.

4.1 Material

We obtained all the controversial topics from ProCon.org (2018). ProCon.org is a non-profit charitable organisation that provides an online resource for search on controversial topics. ProCon.org selects the topics that are controversial and important to many US citizens by also taking the readers' suggestions into account. We collected all 74 controversial topics with their topic questions from the website. Then, we applied three filters on these topics for practical reasons without deliberately selecting any topics. The first filter selects only the *polar* questions, also known as yes-no questions because they have no different sides for the analysis. This filter decreased the topic set size from 74 to 70. The second filter removes the topics that do not contain up-to-date information in their topic pages provided by ProCon.org since they are not *recent* controversial topics and would not return up-to-date results. With the second filter, the number of topics became 64. Lastly, the third filter only includes the topics if both search engines return results for the corresponding topic questions, otherwise the comparison analysis would not be possible. After the last filter, the final topic set became the size of 57. Table 2 contains the full list of controversial topic titles with questions used in this study.

We used the topic questions of these 57 topics for crawling. For example, the topic question of the topic title 'abortion' is 'Should Abortion Be Legal?'. The topic questions reflect the main debate on the corresponding controversial topics and we used them as they are (i.e. including upper-cased characters, without removing punctuation, etc.) for querying the search engines.

We collected the news search results in *incognito mode* to avoid any personalisation effect. Thus, the retrieved SERPs are not specific to anyone, but (presumably) general to US users. We submitted each topic question to US News search engines of Google and Bing using a US proxy. Since we used the *news versions* of the two search engines, sponsoring results which may affect our analysis did not appear in the news search results at all. Then, we firstly crawled the URLs of the retrieved results for the same topic question to minimise the time lags between the search engines since the SERP of the same topic may vary over time. Subsequently, we extracted the textual contents of the top-10 documents using the crawled URLs. By this way, the time span between the SERPs of Google and Bing for each controversial topic (whole corpus) became 2-3 minutes on average. Moreover, before starting the crawling process, we firstly made some experiments with a small set of topics (different from the topic set provided in the paper) in the news search as well as default search and did not observe significant changes especially in the top-10 documents of the news search even in 10-15 minutes time lags. This indicates that the news search is less dynamic than default search and we believe that the 2-3 minutes of time lags would not drastically affect the search results.

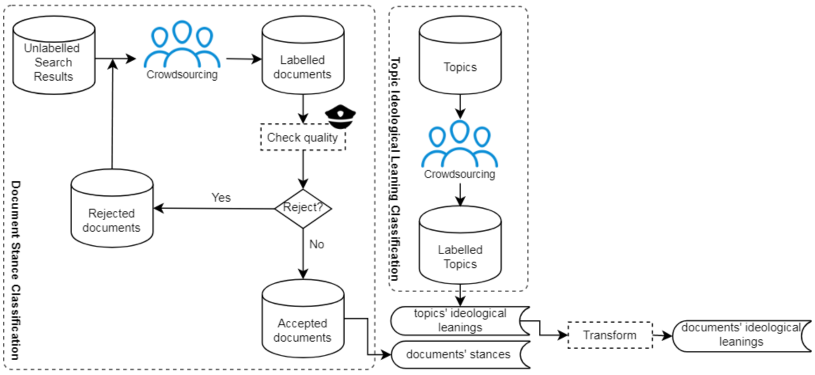

Fig. 1 Flow-chart of the crowd-sourcing campaigns

4.2 Crowd-sourcing Campaigns

The end-to-end process of obtaining stances and ideological leanings is shown in the flow-chart in Figure 1. The emphasised (dotted) parts of the flow-chart show the steps of the Document Stance Classification (DSC) and Topic Ideological Leaning Classification (TILC).

The DSC process inputs unlabelled top-10 search results, crawled by the data collection procedure described in Section 4.1, and outputs the stance labels of all these documents via crowd-sourcing with respect to the topic questions ($\mathcal{Q}$) used to retrieve them. As displayed in the flow-chart, the TILC process uses crowd-sourcing to output the ideological leanings of all topic questions ($\mathcal{Q}$). Then, the accepted stance labels of all documents, acquired from the DSC process are transformed into ideological leaning labels based on the assigned ideology of their corresponding topic questions. The steps of obtaining document labels in stance and ideological leaning detection are described below.

To label the stance of each document with respect to the topic questions ($\mathcal{Q}$) we used crowd-sourcing. We selected MTurk as a crowd-sourcing platform. In this platform, to obtain high quality crowd-labels task properties were set as follows. Since the topics are mostly related to US, we selected crowd-workers only from US. Moreover, we tried to find qualified and experienced workers by setting the following thresholds: Human Intelligence Task (HIT) approval rate percentage should be greater than 95% and number of HITs approved should be greater than 1000 for each worker. We set the wage as 0.15$ and time allowed was 30 minutes per HIT. Each document was judged by three crowd-workers.

To classify the stance of a document we asked crowd-workers to label, given a controversial topic question, the stance of a document in pro, neutral, against, not-relevant, or link not-working. Before the task was assigned, instructions were given to a worker in three groups from general to specific. Initially, workers were provided an overview of the stance detection task, then

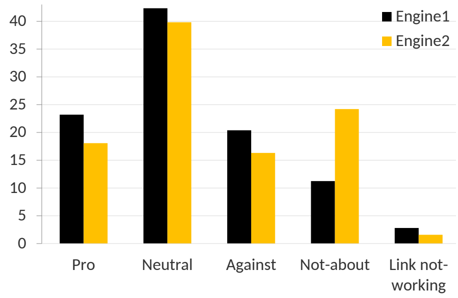

Fig. 2 Percentages of the document stance labels annotated by crowd-workers

steps of the task were listed, i.e. read the topic question, open the news article link etc., and finally, rules and tips were displayed. This last part contained definitions of having a pro, neutral or against stance as given in Section 3.1 above. Additionally, we included a clue for workers saying that title of the article may give you a general idea about the stance, however it is not sufficient to determine its overall viewpoint and then request workers to read also the rest of the article. Apart from these, at the end of the page we put a warning and informed the workers that some of the answers were known to us and we may reject their HITs, i.e. single, self-contained task for a worker, based on evaluation. Then, in the following page a HIT was shown to the worker with a topic question (query), link to the news article whose stance will be determined by repeating/reminding the main question of the stance detection task.

In order to obtain reliable annotations, we first annotated a randomly chosen set of documents later used to check the quality of crowd-labels as specified in the warning to the workers. With these expert labels, we rejected low quality annotations and requested new labels for those documents. This iterative process continued until we obtained all the document labels. At the end of this iterative process, for the sake of label reliability, we computed two agreement scores on the approved labels for document stance detection reported in Table 3. The reported inter-rater agreement scores are the percent agreements between the corresponding annotators. We looked at pairwise agreement; put 1 if there is an agreement and 0, otherwise. Then we computed the mean for the fractions. Reported Kappa score for document stance classification is considered *fair* agreement. Previously, researchers reported a Kappa score of the inter-rater agreement between experts (0.385) instead of crowd-workers for the same task, i.e. document stance classification in SERPs towards a different query set which includes controversial topics as well as popular products, by claiming MTurk workers had difficulty with the task (Alam & Downey, 2014). Although our task seems to be more challenging, i.e. the queries are only about controversial issues, our reported Kappa score for MTurk workers is compara-

Table 3 Crowd-workers Agreement

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

Table 4 Performance of the search engines, p-values of a two-tailed paired t-test computed between engine 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

ble to their expert agreement score, which we believe to be sufficient due to the subjective nature and difficulty of the task.

The distribution of the accepted stance labels for the search results of each search engine is displayed in Figure 2. One may argue that for a query about a controversial topic issued to a news search engine, its SERP would mostly contain controversial articles that support one dominant viewpoint towards a given topic. Hence, informational pages or articles adequately discussing different viewpoints of the topic, i.e. documents that have a neutral stance, would never get a chance to be included in the analysis. However, the distribution in Figure 2 refutes this argument by showing that the majority of the labels for both search engines is actually *neutral*.

To identify the ideological leaning of each topic, we again used crowdsourcing as displayed in Figure 1. We asked the crowd-workers to classify each topic as: conservative, liberal, or both or neither. To get high quality annotations also for topic ideology detection, worker properties were set as the same with the stance detection. We again selected crowd-workers only from US. The wage per HIT was set as 0.1$ and the time allowed was 5 minutes. Similarly to the stance detection, in the informational page we gave an overview, listed the steps and lastly provided the rules & tips. For this task, last part contained the ideological leaning definitions as given in Section 3.1. Additionally, we requested the workers to evaluate the ideological leaning of a given topic based on the current ideological climate and warned them related to the rejection of their HITs as before. In the next page, the workers were shown a HIT with a topic question (query), i.e. one of the main debates of the corresponding topic, and asked the worker the following: *Which ideological group would answer favourably to this question?*. The topics assigned to conservative or liberal leanings have been decided based on the judgment of five annotators with majority-voting. The leanings of the topics are shown in Table 2. Two agreement scores computed on the judgments for ideological leaning detection are also reported in Table 3.

To map the stance from the *pro-to-against* to the *conservative-to-liberal*, we applied a simple transformation to the documents. This transformation is

Table 5 Stance bias of the search engines, p-values of a two-tailed paired t-test computed between engine 1 and 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 6 Ideological bias of the search engines, p-values of a two-tailed paired t-test computed between engine 1 and 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

needed because there may be documents which have a pro stance, for example, towards *abortion* and *Cuba embargo*. Though these documents have the same stance, they have different ideological leanings since having a pro stance on *abortion* implies a *liberal leaning*, whereas a pro stance on *Cuba embargo* implies a *conservative leaning*. For some topics (as in the case of Cuba embargo), we can directly interpret the *pro-to-against* stance labels of search results as *conservative-to-liberal* ideological leaning labels while for other topics (as in the case of the abortion) as *liberal-to-conservative*. On the other hand, for those topics such as *vaccines for kids*, which crowded label resulted in both or neither, the conservative-to-liberal or liberal-to-conservative transformation was not meaningful and therefore eliminated by our analysis. We note that within budget constraints, the crowd-sourcing protocol was designed to obtain crowd-labels with high-quality by labelling (expert) the random sample of documents, applying iterative process and majority voting on these labels.

4.3 Results

In Table 4 we present the performance of the two search engines. This is measured over all the topics. A document is considered relevant when classified as pro, against, or neutral. The difference for all evaluation measures is statistically significant.

In Table 5 we present the stance bias of the search engines. Note that for all the three measures of bias, P@10, RBP and DCG@10, lower value is better which means lower bias in the scope of this work as opposed to their corresponding classic IR measures. All MB and MAB scores are positive

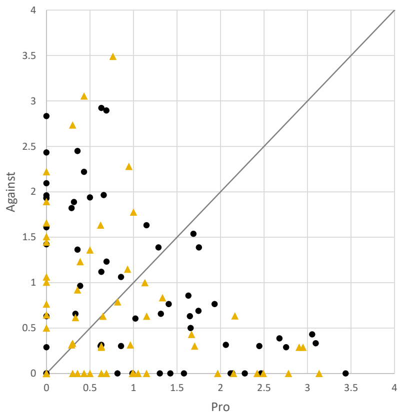

Fig. 3 *DCG*👍@10 against *DCG*👎@10 measured on stances – black points for engine 1 and yellow points for engine 2

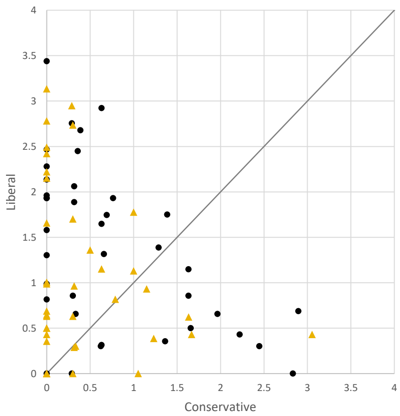

Fig. 4 *DCG*🔴@10 against *DCG*🔵@10 measured on ideological leanings – black points for engine 1 and yellow points for engine 2

for all three IR evaluation measures. Also, the differences between the two search engines for both MB and MAB measures are statistically not significant and it is shown with the two-tailed pair t-test on these measures. In Table 6 we show the ideological bias. Similarly to Table 5, lower is better since we use the same measures of bias. This table is similar to Table 5. Unlike the Table 5, all MB scores are negative while all MAB scores are positive for all three IR evaluation measures. The two-tailed paired t-test computed on MBs to compare the difference in bias between engine 1 and engine 2, this is statistically not significant. Nonetheless, the two-tailed test on MABs is statistically not significant for the measure P@10; but it is statistically significant for the measures RBP and DCG@10.

In Figure 3 we show how the topic-wise SERPs distribute over the pro-against stance space for the measure DCG@10. The x-axis is the pro stance score (*DCG*👍@10) and the y-axis is the against stance score (*DCG*👎@10). Each point corresponds to the overall SERP score of a topic. Black points are those SERPs retrieved by engine 1 and yellow points are those retrieved by engine 2.

In Figure 5 we compare the overall stance bias score ($\beta_{DCG@10}$), i.e. difference between the pro and against stance scores, of SERPs for each topic measured on the two search engines. The x-axis is engine 1 and the y-axis is engine 2. The points in positive coordinates denote the topics whose SERPs are overall biased towards the pro stance, negative coordinates are for the against stance.

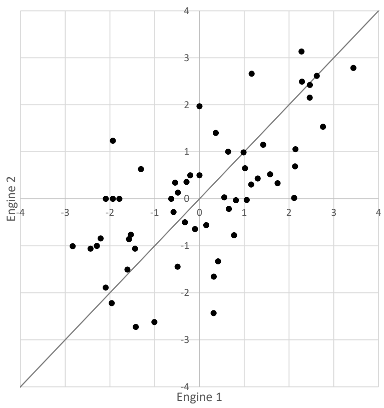

Fig. 5 $\beta_{DCG@10}$ measured on stances, where positive is 👍 and negative is 👎

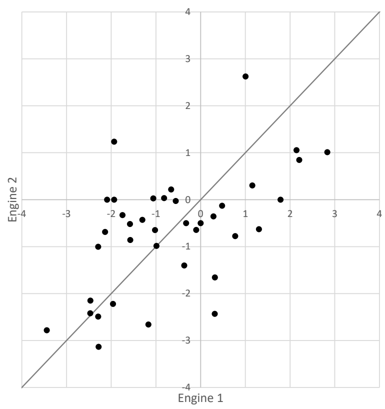

Fig. 6 $\beta_{DCG@10}$ measured on leanings, where positive is 🔴 and negative is 🔵

Figure 4 and Figure 6 are similar to Figure 3 and Figure 5 but instead of measuring the stance bias we measure the ideological bias in the former case. Therefore, Figure 4 displays how the overall SERPs of topics distribute over the conservative-liberal ideological space for the measure DCG@10. Similarly, in Figure 6 we compare the overall ideological bias score ($\beta_{DCG@10}$), i.e. difference between the conservative and liberal leaning scores, of the SERPs where the points in positive coordinates stand for the topics that are biased towards the conservative leaning, negative coordinates are for the liberal.

5 Discussion

Before investigating the existence of bias in SERPs, we initially compared the retrieval performances of two search engines. In Table 4 we observe that the performance of the two search engines is high but engine 1 is better than engine 2 – their difference is statistically significant. This is verified across all three IR evaluation measures.

Next, we verify if the search engines return biased results in terms of document stances (RQ1) and if so, we further investigate if the engines suffer from the same level of bias (RQ2) that the difference between the engines are not statistically significant. In Table 5 all MB scores are positive and regarding the RQ1, the engines seem to be biased towards the pro stance. We applied the one-sample t-test on MB scores to check the existence of stance bias, i.e. if the true mean is different from zero, as mentioned in Section 3.3. However, these biases are statistically not significant which means that this expectation may be the result of noise – there is not a systematic stance bias, i.e. preference of one stance with respect to the other. Based on MAB scores, we can observe that both engines suffer from an absolute bias. However, the difference between the two engines is shown to be non-significant with the two-tailed t-test. These results show that both search engines are not biased towards a *specific* stance in returning results since there is no statistically significant difference from the *ideal* distribution. Nonetheless, for both engines there exists an absolute bias which can be interpreted as the expected bias for a topic question. These empirical findings imply that the search engines are biased for some topics towards the pro stance and for others towards the against stance.

The results are displayed in Figure 3. This figure refers to the values used to compute the MAB score of the DCG@10 column. It shows that the difference between the pro and against stances of both engines for topics is uniformly distributed. To note that, no topic can be located on the up-right area of the plot because the sum of their coordinates is bounded by the maximum possible DCG@10 score. Moreover we observe that topics are distributed similarly across the engines. This is also confirmed by Figure 5 where we can observe that the stance bias scores ($\beta_{DCG@10}$), i.e. the differences between DCG@10 scores for the pro stance and DCG@10 scores for the against stance, of topics are somehow balanced between the up-right quadrant and the low-left quadrant. Moreover, these two quadrants are the area of agreement in stance between the two engines. The other two quadrants contain those topics where the engines disagree. Here we can conclude that the engines agree with each other in the majority of cases.

Lastly, we investigate if the search engines are biased in the ideology space (RQ3). Looking at MB scores in Table 6 we observe that both search engines seem to be biased towards the same ideological leaning – liberal (all MB scores are negative). Unlike the stance bias, one sample t-test on MB scores show that these expectations are statistically significant with different confidence values, i.e. p-value < 0.005 across all three IR measures for engine 2; whereas the same confidence value on P@10 for engine 1 and p-value < 0.05 on RBP and DCG@10. These results indicate that both search engines are biased towards the same leaning which is liberal. Comparing the two search engines on MB scores, we observe that their differences are statistically not significant, which means that the observed difference may be the result of random noise. Based on MAB, since all MAB scores are positive we can also observe that both engines suffer from an absolute bias. However, in contrast with what observed for the stance bias, this time there is a difference in expected ideological bias between the two search engines. For RBP and DCG@10 the difference between the engines is statistically significant. This finding and the different user models that these evaluation measures model suggest that the perceived bias by the users may change based on their behaviour. A user that always inspects the first 10 results (as modelled by P@10) may perceive the same ideological bias between engine 1 and engine 2, while a less systematic user, which just inspects the top results, may perceive that engine 1 is more biased than engine 2. Moreover, comparing this finding with the performance of the engines, we can observe that the better performing engine is more biased than the worse performing one.

Comparing Figure 4 with Figure 3 we observe that in Figure 4 the points look less uniformly distributed than in Figure 3. Topics are mostly on the liberal side. Moreover, engine 2 has fewer points on the conservative side than engine 1. Comparing Figure 6 with Figure 5, we observe that the engines in Figure 6 are more biased towards the liberal side with respect to what observed in Figure 5. Also, we observe that the engines mostly agree – most of the points are placed on the up-right and low-left quadrants.

In conclusion, we find important to point out that it is not in the scope of this work to find the source of bias. As discussed in the introduction, bias may be a result of the input data, which may contain biases, or the search algorithm, which contains sophisticated features and specifically chosen algorithms that, although designed to be effective in satisfying information needs, may produce systematic biases. Nonetheless, we look at the problem from the user perspective and no matter where the bias comes from; the results are biased as described. Our findings seem to be consistent with prior works (Epstein & Robertson, 2017; Diakopoulos *et al.*, 2018) that there exists liberal (left-leaning) partisan bias in SERPs; even in unpersonalised search settings (Robertson *et al.*, 2018a).

6 Limitations

This work has potential limitations. As stated in the introduction, we focus on a particular kind of bias, known as *statistical parity*, or more generally known as *equality of outcome* instead of *equality of opportunity* which uses query-specific base rates. In the context of the controversial topics where the document labels were obtained via crowd-sourcing, this bias measure, i.e. requiring equal representation of stances instead of query-specific base rates, made our experiments feasible. This is firstly because, not all of the query questions in our list have certain answers based on scientific facts, i.e. some of them are subjective queries. In investigating the equality of opportunity, queries can be further categorized as subjective and objective on top of our evaluation framework. For the objective queries, expert labels can be obtained and used as base rates, then search results can be evaluated by taking into account these base rates. Please note that our evaluation framework could better be applied to the controversial queries from the public's perspective mainly where the goal is to have balanced SERPs instead of skewed results. We believe that some queries should be handled with a different framework since those queries are not intrinsically controversial such as *Is Holocaust real?* - there is only one correct answer without the need of a discussion.

Besides, the identification of the stance for the full ranking list is currently too expensive to get annotated via crowd-sourcing. To tackle this issue, a machine learning model can help us to automate the process of obtaining the stance labels. Another potential limitation is that some queries may not be real user queries. Nonetheless, we extracted the queries directly from their topic pages of the ProCon.org (2018) along with the topics. We deliberately did not change the queries to avoid any interference/bias from our side on the results. In this work, we did not make a domain-specific selection of the topics, or apply any filtering as subjective/objective, rather we accepted them as *controversial* topics from the general public's perspective which is the main scope of this work.

Apart from these, crowd-workers' own personal biases may affect the labelling process. For this reason, we tried to mitigate these biases by i. asking the workers to annotate stances rather than ideologies to make their judgment more objective, and ii. aggregating the final judgment coming from multiple workers. Additionally, our analysis refers to a specific point in time where the data was collected. To enable reproducibility and an easier comparison of these results at some point in the future, we made our dataset publicly available. Lastly, we note that this bias analysis can only be used as an indicator of potentially biased ranking algorithms because it is not enough in order to track the source of bias. In the scope of this work, we did not investigate the source of bias that may come from the data (input bias) or from the ranking mechanism (algorithmic bias) of the corresponding search engines. Despite these potential limitations, we believe that our work is a good attempt to evaluate bias in search results with new bias measures and a dataset crawled specifically for the search bias evaluation. Since the bias analysis is very complex, we deliberately limited our scope and only focused on the bias analysis of *recent* controversial topics in *news* search. Nonetheless, all these limitations lead us to numerous interesting future directions.

7 Conclusion & Future Work

In this work we introduced new bias evaluation measures and a generalisable evaluation framework to address the issue of web search bias in news search results. We applied the proposed framework to measure stance and ideological bias in the SERPs of Bing and Google as well as compare their relative bias towards controversial topics. Our 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. In this work, we intended 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.

In the scope of this work we did not investigate the source of bias which we left as future work, therefore the results can be seen as a potential indicator. In our experiments, we gathered document stances via crowd-sourcing. Thus, the obvious future work in this direction is to use automatic stance detection methods instead of crowd-sourcing to obtain the document labels, thereby evaluating bias in the whole corpus of retrieved SERPs to track the source of bias. Moreover, investigating the workers' bias in a follow-up work would be interesting since it is very difficult to remove all biases in practice. In this work, we focus on *equality of outcome*; but using another bias measure, *equality of opportunity* which takes into account the corresponding group proportions, i.e. query-specific base rates, in the population would be an alternative follow-up work. We plan to categorize queries as subjective and objective, then modify the *ideal* ranking definition specifically for the objective queries based on the corpus distributions. The bias analysis for the objective queries, particularly the ones related to the critical domains such as health search, can be investigated further on top of our evaluation framework which we believe to be an interesting follow-up work. Furthermore, we plan to study the effect of localization and personalization, i.e. how much the stances and ideological leanings varied across users or the echo chamber effect, on SERPs, then incorporate that study into our bias evaluation framework in the future.

Compliance with Ethical Standards

Author Emine Yilmaz previously worked as a research consultant for Microsoft Research and she is currently a research consultant for Amazon Research.

Acknowledgements

We thank the reviewers for their comments. This work has been funded by the EPSRC Fellowship titled "Task Based Information Retrieval", grant reference number EP/P024289/1 and the visiting researcher programme of The Alan Turing Institute.

References

(2018). Internetlivestats. `http://www.internetlivestats.com/`, accessed: 2018-10-06.

(2018). Procon.org, procon.org - pros and cons of controversial issues. `https://www.procon.org/`, accessed: 2018-07-31.

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

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. *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. *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**, 1053–1066.

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

Baeza-Yates, R. (2016). Data and algorithmic bias in the web. *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**, 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**, 1014.

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*, .

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

Chelaru, S., Altingovde, I.S. & Siersdorfer, S. (2012). Analyzing the polarity of opinionated queries. *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**, 6.

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

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

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). *ACM SIGIR Forum*, vol. 52, pp. 46–47, ACM New York, NY, USA.

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

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**.

Diaz, A. (2008). Through the google goggles: Sociopolitical bias in search engine design. *Web search*, pp. 11–34, Springer.

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, .

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**, 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.

Fang, Y., Si, L., Somasundaram, N. & Yu, Z. (2012). Mining contrastive opinions on political texts using cross-perspective topic model. *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. *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. *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*, pp. 229–236.

Gao, R. & Shah, C. (2020). Toward creating a fairer ranking in search engine results. *Information Processing & Management*, **57**, 102138.

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

Geyik, S.C., Ambler, S. & Kenthapadi, K. (2019). Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. *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. *Web Search*, pp. 121–133, Springer.

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.

Hu, D., Jiang, S., E. Robertson, R. & Wilson, C. (2019). Auditing the partisanship of google search snippets. *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*, .

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

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**, 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. *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*, pp. 1–40.

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

Lawson, N., Eustice, K., Perkowitz, M. & Yetisgen-Yildiz, M. (2010). Annotating large email datasets for named entity recognition with mechanical turk. *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.

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. *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.

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

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

Mowshowitz, A. & Kawaguchi, A. (2005). Measuring search engine bias. *Information processing & management*, **41**, 1193–1205.

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

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

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

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

Otterbacher, J., Bates, J. & Clough, P. (2017). Competent men and warm women: Gender stereotypes and backlash in image search results. *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. *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**, 801–823.

Räbiger, S., Gezici, G., Saygın, Y. & Spiliopoulou, M. (2018). Predicting worker disagreement for more effective crowd labeling. *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. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 429–435.

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

Robertson, R.E., Lazer, D. & Wilson, C. (2018b). Auditing the personalization and composition of politically-related search engine results pages. *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. *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pp. 1679–1684, ACM.

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. *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.

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. *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. *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 13, pp. 505–516.

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

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

Yang, K. & Stoyanovich, J. (2017). Measuring fairness in ranked outputs. *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*, p. 22, ACM.

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

Yuen, M.C., King, I. & Leung, K.S. (2011). A survey of crowdsourcing systems. *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. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pp. 1569–1578, ACM.