

Document Version

Final published version

Citation (APA)

Draws, T., Roy, N., Inel, O., Rieger, A., Hada, R., Yalcin, M. O., Timmermans, B., & Tintarev, N. (2023). Viewpoint Diversity in Search Results. In J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, A. Caputo, & U. Kruschwitz (Eds.), *Advances in Information Retrieval - 45th European Conference on Information Retrieval, ECIR 2023, Proceedings* (pp. 279-297). (Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Vol. 13980). Springer. https://doi.org/10.1007/978-3-031-28244-7_18

Important note

To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright

In case the licence states "Dutch Copyright Act (Article 25fa)", this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership.
Unless copyright is transferred by contract or statute, it remains with the copyright holder.

Sharing and reuse

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy

Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

Green Open Access added to TU Delft Institutional Repository









'You share, we take care!' - Taverne project

<https://www.openaccess.nl/en/you-share-we-take-care>

Otherwise as indicated in the copyright section: the publisher is the copyright holder of this work and the author uses the Dutch legislation to make this work public.



Viewpoint Diversity in Search Results

Tim Draws¹ , Nirmal Roy¹ , Oana Inel² , Alisa Rieger¹ ,
Rishav Hada³ , Mehmet Orcun Yalcin⁴ , Benjamin Timmermans⁵ ,
and Nava Tintarev⁶ 

¹ Delft University of Technology, Delft, Netherlands

{t.a.draws,n.roy,a.rieger}@tudelft.nl

² University of Zurich, Zurich, Switzerland
inel@ifi.uzh.ch

³ Microsoft Research, Bangalore, India

⁴ Independent Researcher, Istanbul, Türkiye

⁵ IBM, Amsterdam, Netherlands

b.timmermans@nl.ibm.com

⁶ Maastricht University, Maastricht, Netherlands

n.tintarev@maastrichtuniversity.nl

Abstract. Adverse phenomena such as the *search engine manipulation effect* (SEME), where web search users change their attitude on a topic following whatever most highly-ranked search results promote, represent crucial challenges for research and industry. However, the current lack of automatic methods to comprehensively measure or increase viewpoint diversity in search results complicates the understanding and mitigation of such effects. This paper proposes a viewpoint bias metric that evaluates the divergence from a pre-defined scenario of ideal viewpoint diversity considering two essential viewpoint dimensions (i.e., *stance* and *logic of evaluation*). In a case study, we apply this metric to actual search results and find considerable viewpoint bias in search results across queries, topics, and search engines that could lead to adverse effects such as SEME. We subsequently demonstrate that viewpoint diversity in search results can be dramatically increased using existing diversification algorithms. The methods proposed in this paper can assist researchers and practitioners in evaluating and improving viewpoint diversity in search results.

Keywords: Viewpoint diversity · Metric · Evaluation · Bias · Search results

1 Introduction

Web search is increasingly used to inform important personal decisions [16, 31, 45] and users commonly believe that web search results are accurate, trustworthy, and unbiased [53]. However, especially for search results related to debated topics, this perception may often be false [30, 54, 65, 66]. Recent research has demonstrated that a lack of viewpoint diversity in search results can lead to undesired outcomes such as the *search engine manipulation effect* (SEME), which occurs when users change their attitude on a topic following whichever viewpoint

happens to be predominant in highly-ranked search results [5, 6, 24, 27, 52]. For instance, SEME can lead users to judge medical treatments as (in-)effective [52] or prefer a particular political candidate over another [27]. To mitigate potential large-scale negative consequences of SEME for individuals, businesses, and society, it is essential to evaluate and foster viewpoint diversity in search results.

Measuring and increasing the diversity of search results has been studied extensively in recent years, e.g., to satisfy pluralities of search intents [2, 19, 58] or ensure fairness towards protected classes [9, 70, 72, 73]. First attempts in specifically evaluating [23, 43] and fostering [47, 61] *viewpoint diversity* in ranked outputs have also been made. However, two essential aspects have not been sufficiently addressed yet: (1) current methods only allow for limited viewpoint representations (i.e., one-dimensional, often binary) and (2) there is no clear conceptualization of viewpoint diversity or what constitutes viewpoint bias in search results. Current methods often assume that any top k portion of a ranked list should represent all available (viewpoint) categories proportionally to their overall distribution, i.e., analogous to the notion of *statistical parity* [23, 42], without considering other notions of diversity [64]. This impedes efforts to meaningfully assess viewpoint bias in search results or measure improvements made by diversification algorithms. We thus focus on three research questions:

- RQ1.** What metric can thoroughly measure viewpoint diversity in search results?
- RQ2.** What is the degree of viewpoint diversity in actual search results?
- RQ3.** What method can foster viewpoint diversity in search results?

We address **RQ1** by proposing a metric that evaluates viewpoint bias (i.e., deviation from viewpoint diversity) in ranked lists using a two-dimensional viewpoint representation developed for human information interaction (Sect. 3). We show that this metric assesses viewpoint diversity in a more comprehensive fashion than current methods and apply it in a case study of search results from two popular search engines (**RQ2**; Sect. 4). We find notable differences in search result viewpoint diversity between queries, topics, and search engines and show that applying existing diversification methods can starkly increase viewpoint diversity (**RQ3**; Sect. 4.3). All code and data are available at <https://osf.io/kz3je/>.

2 Related Work

Viewpoint Representations. *Viewpoints*, sometimes called *arguments* [3, 26] or *stances* [41], are positions or opinions concerning debated topics or claims [20]. To *represent* viewpoints in ranked lists of search results, each document needs to receive a label capturing the viewpoint(s) it expresses. Previous work has predominantly assigned binary (e.g., *con/pro*) or ternary (e.g., *against/neutral/in favor*) viewpoint labels [32, 52, 71]. However, these labels ignore the viewpoint’s degree and reason behind opposing or supporting a given topic [20], e.g., two statements *in favor* of school uniforms could express entirely different viewpoints in strongly supporting school uniforms for productivity reasons and only

somewhat supporting them for popularity reasons. To overcome these limitations, earlier work has represented viewpoints on ordinal scales [23, 24, 43, 57], continuous scales [43], as multi-categorical perspectives [3, 18, 21], or computed the viewpoint *distance* between documents [47]. A recently proposed, more comprehensive viewpoint label [20], based on work in the communication sciences [7, 8, 11], consists of two dimensions: *stance* (i.e., an ordinal scale ranging from “strongly opposing” to “strongly supporting”) and *logics of evaluation* (i.e., underlying reasons –sometimes called *perspectives* [18, 21], *premises* [13, 26] or *frames* [3, 47]).

Viewpoint Diversity in Ranked Outputs. Previous research has shown that search results across topics and domains (e.g., politics [54], health [65, 66]) may not always be viewpoint-diverse and that highly-ranked search results are often unbalanced concerning query subtopics [30, 50]. Limited diversity, or bias, can root in the overall search result index but become amplified by biased queries and rankings [30, 56, 66]. Extensive research further shows that viewpoint-biased (i.e., *unbalanced*) search results can lead to undesired consequences for individuals, businesses, and society (e.g., SEME) [5, 10, 24, 27, 52, 67]. That is why many studies now focus on understanding and mitigating cognitive user biases in this context [6, 24, 28, 33, 44, 51, 57, 68, 69, 71]. However, because adverse effects in web search are typically an interplay of content and user biases [67], it is essential to also develop methods to evaluate and foster viewpoint diversity in search results.

Building on work that measured diversity or fairness in search results concerning more general subtopics [2, 9, 19, 70, 72, 73], recent research has begun to evaluate *viewpoint diversity* in ranked outputs. Various metrics have been adapted from existing information retrieval (IR) practices to quantitatively evaluate democratic notions of diversity [37, 63, 64], though only few [63] crucially incorporate users’ attention drop over the ranks [6, 27, 39, 49]. *Ranking fairness metrics* such as *normalized discounted difference* (rND) [70] can assess viewpoint diversity by measuring the degree to which documents of a pre-defined protected viewpoint category are ranked lower than others [23]. The recently proposed *ranking bias* (RB) metric considers the full range of a continuous viewpoint dimension and evaluates viewpoint balance [43]. Existing metrics such as rND and RB, however, have a key limitation when measuring viewpoint diversity: they cannot accommodate comprehensive, multi-dimensional viewpoint representations. Incorporating such more comprehensive viewpoint labels is crucial because stances and the reasons behind them can otherwise not be considered simultaneously [20].

Search Result Diversification. To improve viewpoint diversity, we build on earlier work on diversifying search results concerning *user intents* [1, 2, 25, 40, 59]. *xQuAD* [59] and *HxQuAD* [38] are two such models that re-rank search results with the aim of fulfilling diverse ranges of information needs at high ranks. Whereas xQuAD diversifies for single dimensions of (multi-categorical) subtopics, HxQuAD adapts xQuAD to accommodate multiple dimensions of subtopics and diversifies in a multi-level hierarchical fashion. For example, for the query *java*, two first-level subtopics may be *java island* and *java programming*.

For the former, queries such as *java island restaurant* and *java island beach* may then be second-level subtopics. To the best of our knowledge, such methods have so far not been used to foster viewpoint diversity in ranked lists.

3 Evaluating Viewpoint Diversity in Search Results

This section introduces a novel metric for assessing viewpoint diversity in ranked lists such as search results. To comprehensively capture documents' viewpoints, we adopt the two-dimensional viewpoint representation recently introduced by Draws et al. [20] (see Sect. 2). Each document thus receives a single *stance* label on a seven-point ordinal scale from strongly opposing (-3) to strongly supporting (3) a topic and anywhere from no to seven *logic of evaluation* labels that reflect the underlying reason(s) behind the stance (i.e., *inspired*, *popular*, *moral*, *civic*, *economic*, *functional*, *ecological*). Although other viewpoint diversity representations could be modeled, this 2D representation supports more nuanced viewpoint diversity analyses than current approaches, and it is still computationally tractable (i.e., only seven topic-independent categories per dimension).

We consider a set of documents retrieved in response to a query (e.g., “**school uniforms well-being**”) related to a particular debated topic (e.g., mandatory school uniforms). R is a ranked list of N retrieved documents (i.e., by the search engine), $R_{1\dots k}$ is the top- k portion of R , and R_k refers to the k^{th} -ranked document. We refer to the sets of stance and logic labels of the documents in R as \mathcal{S} and \mathcal{L} , respectively, and use \mathcal{S}_k or \mathcal{L}_k to refer to the labels of the particular document at rank k . For instance, a document at rank k may receive the label [$\mathcal{S}_k = 2$; $\mathcal{L}_k = (\textit{popular}, \textit{functional})$] if the article **supports** (stance) school uniforms because they supposedly are popular among students (i.e., **popular** logic) and lead to better grades (i.e., **functional** logic). S and L , respectively, are the (multinomial) stance and logic distributions of the documents in R .

Defining Viewpoint Diversity. Undesired effects such as SEME typically occur when search result lists are one-sided and unbalanced in terms of viewpoints [6, 27, 52]. To overcome this, we follow the normative values of a *deliberative democracy* [37], and counteract these problems through viewpoint plurality and balance. We put these notions into practice by following three intuitions:

1. *Neutrality.* A set of documents should feature both sides of a debate equally and not take any particular side when aggregated. We consider a search result list as neutral if averaging its stance labels results in 0 (a neutral stance score).
2. *Stance Diversity.* A set of documents should have a balanced stance distribution so that different stance strengths (e.g., 1, 2, and 3) are covered. For example, we consider a search result list as stance-diverse if it contains equal proportions of all seven different stance categories, but not if it contains only the stance categories -3 and 3 (albeit satisfying *neutrality* here).
3. *Logic Diversity.* A set of documents should include a plurality of reasons for different stances (i.e., balanced logic distribution *within* each stance category). For example, a search result list may not satisfy *logic diversity* if documents containing few reasons (here, logics) are over-represented.

Our metric *normalized discounted viewpoint bias* (nDVB) measures the degree to which a ranked list *diverges* from a pre-defined scenario of ideal viewpoint diversity. It combines the three sub-metrics *normalized discounted polarity bias* (nDPN), *normalized discounted stance bias* (nDSB), and *normalized discounted logic bias* (nDLB), which respectively assess the three characteristics of a viewpoint-diverse search result list (i.e., *neutrality*, *stance diversity*, and *logic diversity*).

3.1 Measuring Polarity, Stance, and Logic Bias

We propose three sub-metrics that contribute to nDVB by considering different document aspects. They all ignore irrelevant during their computation and – like other IR evaluation metrics [55] – apply a discount factor for rank-awareness.

Normalized Discounted Polarity Bias (nDPB). Polarity bias considers the mean stance label balance. *Neutrality*, the first trait in our viewpoint diversity notion, posits that the stance labels for documents in any top k portion should balance each other out (mean stance = 0). We assess how much a top k search result list *diverges* from this ideal scenario (i.e., *polarity bias*; PB; see Eq. 1) by computing the average normalized stance label. Here, $S_{1\dots k}$ is the set of stance labels for all documents in the top k portion of the ranking. PB normalizes all stance labels S_i in the top k to a score between -1 and 1 (by dividing it by its absolute maximum, i.e., 3) and takes their average. To evaluate the neutrality of an entire search result list τ with N documents, we compute PB iteratively for the top $1, 2, \dots, N$ ranking portions, aggregate the results in a discounted fashion, and apply min-max normalization to produce nDPB (see Eq. 2). Here, Z is a normalizer equal to the highest possible value for the aggregated and discounted absolute PB values and I is an indicator variable equal to -1 if $\sum_{k=1}^N \frac{\text{PB}(S, k)}{\log_2(k+1)} < 0$ and 1 otherwise. nDPB quantifies a search result list’s bias toward opposing or supporting a topic and ranges from -1 to 1 (more extreme values indicate greater bias, values closer to 0 indicate neutrality).

$$\text{PB}(S, k) = \frac{\sum_{i=1}^k \frac{S_i}{3}}{|S_{1\dots k}|} \quad (1) \quad \text{nDPB}(\tau) = \frac{1}{Z} I \sum_{k=1}^N \frac{|\text{PB}(S, k)|}{\log_2(k+1)} \quad (2)$$

Normalized Discounted Stance Bias (nDSB). Stance bias evaluates how much the stance distribution diverges from the viewpoint-diverse scenario. *Stance diversity*, the second trait of our viewpoint diversity notion, suggests that all stance categories are equally covered in any top k ranked list portion. We capture this ideal scenario of a balanced stance distribution in the uniform target distribution $T = (\frac{1}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7})$. The stance distribution of the top k -ranked documents is given by $S_{1\dots k} = (\frac{|S_{1\dots k}^{-3}|}{k}, \dots, \frac{|S_{1\dots k}^3|}{k})$, where each numerator refers to the number of top- k search results in a stance category. We assess how much $S_{1\dots k}$ diverges from T by computing their *Jensen-Shannon divergence* (JSD), a symmetric distance metric for discrete probability distributions [29]. This approach is inspired by work suggesting divergence metrics to measure viewpoint diversity [23, 63, 64]. We then normalize JSD between $S_{1\dots k}$ and T by dividing

it by the maximal divergence, i.e., $JSD(U||T)$ where $U = (1, 0, 0, 0, 0, 0, 0)$ and call the result *stance bias* (SB; see Eq. 3). SB ranges from 0 (desired scenario of stance diversity) to 1 (maximal stance bias). Notably, SB will deliberately *always* return high values for the very top portions (e.g., top one or two) of any search result list, as it is impossible to get a balanced distribution of the seven stance categories in just a few documents. We evaluate an entire search result list using nDSB (see Eq. 4), by computing SB iteratively for the top $1, 2, \dots, N$ ranking portions, aggregating the results in a discounted fashion, and normalizing.

$$SB(S, k) = \frac{JSD(S_{1\dots k}||T)}{JSD(U||T)} \quad (3) \quad nDSB(\tau) = \frac{1}{Z} \sum_{k=1}^N \frac{SB(S, k)}{\log_2(k+1)} \quad (4)$$

Normalized Discounted Logic Bias (nDLB). Logic bias measures how balanced documents in each stance category are in terms of logics. *Logic diversity* suggests that all logics are equally covered in each document group when splitting documents by stance category. Thus, when a search result list contains documents, e.g., with stances $-1, 0,$ and $1,$ the logic distributions of each of those three groups should be balanced. The logic distribution of all top k results belonging to a particular stance category s is given by $L_{1\dots k}^s = \left(\frac{|\mathcal{L}_{1\dots k}^{s,l_1}|}{|\mathcal{L}_{1\dots k}^s|}, \dots, \frac{|\mathcal{L}_{1\dots k}^{s,l_\tau}|}{|\mathcal{L}_{1\dots k}^s|} \right)$, where each numerator $|\mathcal{L}_{1\dots k}^{s,l}|$ refers to the number of times logic l (e.g., *inspired*) appears in the top k documents with stance category s . Each denominator $|\mathcal{L}_{1\dots k}^s|$ is the total number of logics that appear in the top k documents with stance category s . $L_{1\dots k}^s$ reflects the relative frequency of each logic in the top k documents in a specific stance category. Similar to SB, we evaluate the degree to which $L_{1\dots k}^s$ diverges from T by computing the normalized JSD for the logic distributions of each available stance category and then produce *logic bias* (LB) by averaging the results (Eq. 5). Here, \mathcal{S}_k^* is the set of unique stance categories among the top k -ranked documents. LB thus quantifies, on a scale from 0 to 1, the average degree to which the logic distributions diverge from the ideal, viewpoint-diverse scenario where all logics are equally present within each stance category. We produce nDLB by computing LB iteratively for the top $1, 2, \dots, N$ documents and applying our discounted aggregation and normalization procedures (Eq. 6).

$$LB(\mathcal{S}, L, k) = \frac{1}{|\mathcal{S}_k^*|} \sum_{s \in \mathcal{S}_k^*} \frac{JSD(L_{1\dots k}^s||T)}{JSD(U||T)} \quad (5) \quad nDLB(\tau) = \frac{1}{Z} \sum_{k=1}^N \frac{LB(\mathcal{S}, L, k)}{\log_2(k+1)} \quad (6)$$

3.2 Normalized Discounted Viewpoint Bias

To evaluate overall viewpoint diversity, we combine nDPB, nDSB, and nDLB into a single metric, called *normalized discounted viewpoint bias* (nDVB):

$$nDVB(\tau) = I \frac{\alpha |nDPB(\tau)| + \beta nDSB(\tau) + \gamma nDLB(\tau)}{\alpha + \beta + \gamma}$$

Here, I is an indicator variable that equals -1 when $nDPB(\tau) < 0$ and 1 otherwise. The parameters α , β , and γ are weights that control the relative importance of the three sub-metrics. Thus, $nDVB$ measures the degree to which a ranked list of documents diverges from an ideal, viewpoint-diverse scenario. It ranges from -1 to 1 , indicating the direction and severity with which such a ranked list (e.g., search results) is biased (values closer to 0 imply greater viewpoint diversity).

Our proposed metric $nDVB$ allows for a more comprehensive assessment of viewpoint diversity in search results compared to metrics such as rND or RB . It does so by allowing for comprehensive viewpoint representations of search results, simultaneously considering *neutrality*, *stance diversity*, and *logic diversity*.

4 Case Study: Evaluating, Fostering Viewpoint Diversity

This section presents a case study in which we show how to practically apply the viewpoint bias metric we propose ($nDVB$; see Sect. 3.2) and examine the viewpoint diversity of real search results from commonly used search engines, using relevant queries for currently debated topics (i.e., *atheism*, *school uniforms*, and *intellectual property*). Finally, we demonstrate how viewpoint diversity in search results can be enhanced using existing diversification algorithms. More details on the materials and results (incl. figures) are available in our repository.

4.1 Materials

Topics. We aimed to include in our case study three topics that (1) are not scientifically answerable (i.e., with legitimate arguments in both the opposing and supporting directions) and (2) cover a broad range of search outcomes (i.e., consequences for the individual user, a business, or society). To find such topics, we considered the *IBM-ArgQ-Rank-30kArgs* data set [35], which contains arguments on controversial issues. The three topics we (manually) selected from this data set were *atheism* (where attitude change may primarily affect the user themselves, e.g., they become an atheist), *intellectual property rights* (where attitude change may affect a business, e.g., the user decides to capitalize on intellectual property they own), and *school uniforms* (where attitude change may affect society, e.g., the user votes to abolish school uniforms in their municipality).

Queries. We conducted a user study (approved by a research ethics committee) to find, per topic, five different queries that users might enter into a web search engine if they were wondering whether one should be an atheist (individual use case), intellectual property rights should exist (business use case), or students should have to wear school uniforms (societal use case). In a survey, we asked participants to imagine the three search scenarios and select, for each, three “neutral” and four “biased” queries from a pre-defined list. The neutral queries did not specify a particular debate side (e.g., `school uniforms opinions`), while the biased queries prompted opposing (e.g., `school uniforms disadvantages`) or supporting results (e.g., `school uniforms pros`).

We recruited 100 participants from *Prolific* (<https://prolific.co>) who completed our survey for a reward of \$0.75 (i.e., \$8.09 per hour). All participants were fluent English speakers older than 18. For our analysis, we excluded data from two participants who had failed at least one of two attention checks. The remaining 98 participants were gender-balanced (49% female, 50% male, 1% non-binary) and rather young (50% were between 18 and 24). We selected five queries per topic: the three most commonly selected neutral queries and the single most commonly selected opposing- and supporting-biased queries (see Table 1).¹

Table 1. Viewpoint diversity evaluation for all 30 search result lists from Engine 1 and 2: rND, RB, and nDVB (incl. its sub-metrics DPB, DSB, and DLB). Queries were designed to retrieve neutral (neu), opposing (opp), or supporting (sup) results (\leftrightarrow).

Query	\leftrightarrow	Engine 1						Engine 2					
		rND	RB	nDPB	nDSB	nDLB	nDVB	rND	RB	nDPB	nDSB	nDLB	nDVB
why people become atheists or theists	neu	.70	.27	.32	.33	.38	.34	.69	.14	.21	.36	.33	.30
should I be atheist or theist	neu	.68	.13	.24	.39	.44	.35	.80	.04	.05	.51	.40	.32
atheism vs theism	neu	.58	-.06	-.07	.52	.37	-.32	.77	.01	.03	.53	.39	.32
why theism is better than atheism	opp	.47	.19	.22	.28	.35	.29	.53	-.04	-.15	.45	.30	-.30
why atheism is better than theism	sup	.35	.05	.15	.23	.43	.27	.68	.10	.15	.45	.34	.31
why companies maintain or give away IPRs	neu	.77	.46	.49	.41	.45	.45	.97	.61	.60	.48	.51	.53
should we have IPRs or not	neu	.80	.34	.34	.35	.33	.34	.93	.47	.44	.42	.41	.43
IPRs vs open source	neu	.80	.10	.09	.45	.43	.32	.92	.18	.19	.57	.53	.43
why IPRs don't work	opp	.69	.30	.33	.42	.40	.38	.54	.18	.19	.40	.35	.31
should we respect IPRs	sup	.90	.48	.49	.41	.36	.42	.95	.60	.59	.50	.35	.48
why countries adopt or ban school unif.	neu	.59	-.01	.14	.37	.25	.26	.54	-.10	-.11	.37	.20	-.23
should students wear school unif. or not	neu	.62	-.10	-.10	.45	.20	-.25	.85	.14	.15	.42	.19	.26
school unif. well-being	neu	.55	.07	.09	.28	.25	.21	.54	.13	.23	.31	.35	.30
why school unif. don't work	opp	.30	-.22	-.31	.33	.18	-.27	.59	-.01	-.03	.37	.21	-.20
why school unif. work	sup	.89	.43	.49	.38	.27	.38	.92	.45	.03	.50	.39	.36
Overall mean absolute bias		.65	.21	.26	.37	.34	.32	.75	.21	.24	.44	.34	.34

Note. In contrast to the actual queries, we here abbreviate *intellectual property rights* (IPRs) and *uniforms* (unif.).

Search Results. We retrieved the top 50-ranked search results for each of the $3 \times 5 = 15$ queries listed in Table 1 from two of the most commonly used search engines, through web crawling or an API.² This resulted in a data set of $15 \times 2 \times 50 = 1500$ search results, 25 of which (mostly the last one or two results) were not successfully retrieved. The remaining 1475 (i.e., 973 unique) search results were recorded, including their query, URL, title, and snippet.

Viewpoint Annotations. To assign each search result the 2D (stance, logic) viewpoint label (see Sect. 3), we employed six experts, familiar with the three topics, the annotation task, and the viewpoint labels. This is more than the one to three annotators typically employed for IR annotation practices [34, 62]. The viewpoint label consists of *stance* (i.e., position on the debated topic on an ordinal scale ranging from -3 ; strongly opposing; to 3 ; strongly supporting)

¹ Due to error, we used the 2nd most common supporting query for the *IPR* topic.

² The retrieval took place on December 12th, 2021 in the Netherlands.

and *logics of evaluation* (i.e., motivations behind the stance).³ First, the experts discussed annotation guidelines and examples before individually annotating the same set of 30 search results (i.e., two results randomly chosen per query). Then, they discussed their disagreements, created an improved, more consistent set of annotation guidelines, and revised their annotations. Following discussions, their overall agreement increased to satisfactory levels for stance (Krippendorff's $\alpha = .90$) and the seven logics ($\alpha = \{.79, .66, .73, .86, .77, .36, .57\}$). Such agreement values represent common ground in the communication sciences, where, e.g., two trained annotators got $\alpha = \{.21, .58\}$ when annotating *morality* and *economical* frames in news [15]. Each expert finally annotated an equal and topic-balanced share of the remaining 943 unique search results.

4.2 Viewpoint Diversity Evaluation Results

We conducted viewpoint diversity analyses per topic, search engine, and query. Specifically, we examined the overall viewpoint distributions and then measured viewpoint bias in each of the ($15 \times 2 =$) 30 different top 50 search result lists retrieved from the two search engines, by computing the existing metrics rND and RB (see Sect. 2) and our proposed metric incl. its sub-metrics (see Sect. 3).

Overall Viewpoint Distributions. Among the 973 unique URLs in our search results data set, 306, 334, and 263 respectively related to the topics *atheism*, *intellectual property rights* (IPRs), and *school uniforms*. A total of 70 unique search results were judged irrelevant to their topic and excluded from the analysis. Search Engine 1 (SE₁) provided a somewhat greater proportion of unique results for the 15 queries (77%) than Search Engine 2 (SE₂, 69%). For all three topics, supporting stances were more common. Regarding logics, the *school uniforms* topic was overall considerably more balanced than the others. Atheism-related documents often focused on *inspired*, *moral*, and *functional* logics (e.g., religious people have higher moral standards, atheism explains the world better). Documents related to IPRs often referred to *civic*, *economic*, and *functional* logics (e.g., IPRs are an important legal concept, IPRs harm the economy).

Viewpoint Diversity per Query, Topic, and Search Engine. We analyzed the viewpoint diversity of search results using the existing metrics rND, RB, and our proposed (combined) metric nDVB. We slightly adapted rND and RB to make their outcomes better comparable; aggregating both in steps of one and measuring viewpoint *imbalance* (or bias) rather than ranking fairness. Our rND implementation considered all documents with negative stance labels as *protected*, all documents with positive stance labels as *non-protected*, and ignored neutral documents. Computing RB required standardizing all stance labels to scores ranging from -1 to 1 . To compute nDVB, we set the parameters to $\alpha = \beta = \gamma = 1$, i.e., giving all sub-metrics equal weights. Table 1 shows the evaluation

³ Note that viewpoint labels do not refer to specific web search queries, but always to the topic (or claim) at hand. For example, a search result supporting the idea that students should have to wear school uniforms always receives a positive stance label (i.e., 1, 2, or 3), no matter what query was used to retrieve it.

results for all metrics across the 30 different search result lists from the two search engines. Scores closer to 0 suggest greater diversity (i.e., less distance to the ideal scenario), whereas scores further away from 0 suggest greater bias.

Neutrality. As we note in Sect. 3, viewpoint-diverse search result lists should feature both sides of debates equally. While rND does not indicate whether a search result list is biased against or in favor of the protected group [23], the RB and nDPB outcomes suggest that most of the search result lists we analyzed are biased towards *supporting* viewpoints. We observed that results on IPRs tended to be more biased than results on the other topics but, interestingly, we did not observe clear differences between query types. Moreover, except for the *school uniforms* topic, supposedly neutral queries generally returned results that were just as biased as queries targeted specifically at opposing or supporting results.

Stance Diversity. Another trait of viewpoint-diverse search result lists is a balanced stance distribution. Since rND, RB, and nDPB cannot clarify whether all stances (i.e., all categories ranging from -3 to 3) are uniformly represented, we here only inspect the nDSB outcomes. While we did not observe a noteworthy difference between topics or queries, we found that SE₂ returned somewhat more biased results than SE₁. Closer examination of queries where the two engines differed most in terms of nDSB (e.g., *why theism is better than atheism*) revealed that SE₂ was biased in the sense that it often returned fewer opinionated (and more neutral) results than SE₁. Regarding their balance between mildly and extremely opinionated results, both engines behaved similarly.

Logic Diversity. The final characteristic of viewpoint-diverse search result lists concerns their distribution of logics, i.e., the diversity of reasons brought forward to oppose or support topics. When inspecting the nDLB outcomes, we found that logic distributions in the search result lists were overall more balanced than stance distributions (see nDSB results) and similar across search engines and queries. However, we did observe that nDLB on the *school uniforms* topic tended to be lower than for other topics, suggesting that greater diversities of reasons opposing or supporting school uniforms were brought forward.

Overall Viewpoint Diversity. To evaluate overall viewpoint diversity in the search result lists, we examined nDVB, the only metric that simultaneously evaluates divergence from neutrality, stance diversity, and logic diversity. Bias *magnitude* per nDVB ranged from .20 to .53 across results from search engines, with only four out of 30 search result lists being biased against the topic. Regarding topics, search results for neutral queries were somewhat less biased on *school uniforms* compared to *atheism* or *intellectual property rights*.

Interestingly, search results for neutral queries on all topics were often just as viewpoint-biased as those from directed queries. Some queries returned search results with different bias magnitudes (e.g., *school uniforms well-being*) or bias directions (e.g., *atheism vs theism*) depending on the search engine. Moreover, whereas search results for supporting-biased queries were indeed always biased in the supporting direction (i.e., positive nDVB score), results for opposing-biased queries were often also biased towards supporting viewpoints. Figure 1 shows, per topic and search engine, how the absolute nDVB developed on average when

evaluated at each rank. It illustrates that nDVB tended to decrease over the ranks across engines, topics, and queries but highlights that the top, say 10, search results that users typically examine are often much more viewpoint-biased than even the top 30 (i.e., more search results could offer more viewpoints).

4.3 Viewpoint Diversification

We implemented four diversification algorithms to foster viewpoint diversity in search results by (1) re-ranking and (2) creating viewpoint-diverse top 50 search result lists using all unique results from each topic. Specifically, we performed *ternary stance diversification*, *seven-point stance diversification*, *logic diversification* (all based on xQuAD; i.e., diversifying search results according to stance labels in the common ternary format, the seven-point ordinal format, or logic labels, respectively), and *hierarchical viewpoint diversification* (based on HxQuAD; i.e., diversifying search results hierarchically: first for seven-point ordinal stance labels and then, within each stance category, for logic labels; giving both dimensions equal weights). We evaluated the resulting search result lists using nDVB.

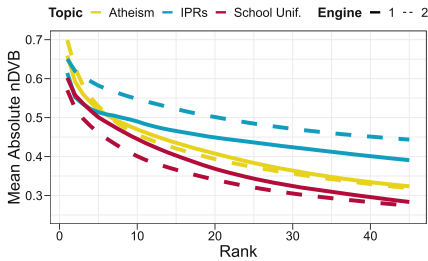


Fig. 1. Development of mean absolute nDVB@ k across search result ranks, split by topic and search engine.

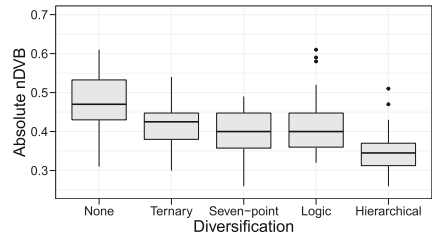


Fig. 2. Mean absolute viewpoint diversity (nDVB@10) per diversification algorithm across the 30 search result lists.

Re-ranked Top 50 Search Result Lists. Fig. 2 compares absolute nDVB between the original top 50 search result lists and the four diversification strategies. All strategies improved the viewpoint diversity of our lists. Whereas the ternary stance diversification only showed marginal improvements (mean abs. nDVB@10 = .42, nDVB@50 = .35) compared to the original search result lists (mean abs. nDVB@10 = .47, nDVB@50 = .33), the hierarchical viewpoint diversification based on stances and logics was the most effective in fostering viewpoint diversity (mean abs. nDVB@10 = .35, nDVB@50 = .27). Viewpoint diversity for the seven-point stance diversification (mean abs. nDVB@10 = .39, nDVB@50 = .29) and logic diversification (mean abs. nDVB@10 = .42, nDVB@50 = .31) were comparable, and in between the ternary stance and hierarchical diversification.

“Best-case” Comparison. Despite the promising re-ranking results, diversification methods can only work with the specific sets of documents they are given. To show a “best-case” scenario for comparison, we employed our diversification algorithms to create, per topic, one maximally viewpoint-diverse search result list using all topic-relevant search results (i.e., from across queries and search engines). We found that all four diversification algorithms yielded search result lists with much less bias when given more documents compared to when they only re-ranked top 50 search results lists. Here, the hierarchical diversification was again most effective (mean abs. nDVB@10 = .29, nDVB@50 = .20); improving by a magnitude of .07 on average over the re-ranked top 50 search result lists. Compared to the average search result list we had retrieved from the two search engines, the “best-case” hierarchical diversification improved viewpoint diversity by margins of .17 (nDVB@10) and .13 (nDVB@50), reflecting a mean improvement of 39%. The other diversification algorithms showed similar improvements, albeit not as impactful as the hierarchical method (i.e., mean abs. nDVB@10 was .37, .37, .34 and mean abs. nDVB@50 was .31, .24, .24 for the ternary stance, seven-point stance, and logic diversifications, respectively).

5 Discussion

We identified that viewpoint diversity in search results can be conceptualized based on the deliberative notion of diversity by looking at *neutrality*, *stance diversity*, and *logics diversity*. Although we were able to adapt existing metrics to partly assess these aspects, a novel metric was needed to comprehensively measure viewpoint diversity in search results. We thus proposed the metric *normalized discounted viewpoint bias* (nDVB), which considers two important viewpoint dimensions (*stances* and *logics*) and measures viewpoint bias, i.e., the deviation of a search result list from an ideal, viewpoint-diverse scenario (RQ1). Findings from our case study suggest that nDVB is sensitive to expected data properties, such as aligning with the query polarity and bias decreasing for larger lists of search results. Although further refinement and investigation of the metric are required (e.g., to find the most practical and suitable balance between the three notions of diversity or outline interpretation guidelines), our results indicate that the metric is a good foundation for measuring viewpoint diversity.

The degree of viewpoint diversity across search engines in our case study was comparable: neither engine was consistently more biased than the other (RQ2). However, we found notable differences in bias magnitude and even bias direction between search engines *regarding the same query* and queries related to the same topic. This lends credibility to the idea that nDVB indeed measures viewpoint diversity, and is able to detect different kinds of biases. Further work is required to compare different metrics and types of biases. Similar to previous research [65], we found that search results were mostly biased in the *supporting* direction. This suggests that actual search results on debated topics may often not reflect a satisfactory degree of viewpoint diversity and instead be systemically biased in terms of viewpoints. More worryingly, depending on where (which search

engine) or how (which query) users search for information, they may not only be exposed to different viewpoints, but ones representing a different bias than their peers. We also found that neutrally formulated queries often returned similarly biased search results as queries calling for specific viewpoints. In light of findings surrounding SEME and similar effects, this could have serious ramifications for individual users' well-being, business decision-making, and societal polarization.

Our case study further showed that diversification approaches based on xQuAD and HxQuAD can improve the viewpoint diversity in search results. Here, the hierarchical viewpoint diversification (based on HxQuAD, and able to consider both documents' *stances* and *logics of evaluation*) was most effective (**RQ3**).

Limitations and Future Work. Although our case study covered debated topics with consequences for individuals, businesses, and society, it is important to note that our results may not generalize to all search engines and controversial issues. We carefully selected the deliberative notion of diversity to guide our work as we believe it suits many debated topics, especially those with legitimate arguments on all sides of the viewpoint spectrum. However, we note that some scenarios may require applying other diversity notions and that presenting search results according to the deliberative notion of diversity (i.e., representing all viewpoints equally) may even cause harm to individual users or help spread fake news (e.g., considering health-related topics where only one viewpoint represents the scientifically correct answer [5, 10, 52, 67]). Future work could measure search result viewpoint bias for larger ranges of topics, explore whether different diversity notions apply when debated topics have clear scientific answers [14, 48, 63], and capture user perceptions of diversity [36, 46, 57].

Another limitation of our work is that, despite providing a diverse range of queries to choose from, queries may not have represented all users adequately. Future work could collect topics and queries via open text fields [67]. Furthermore, our proposed metric nDVB is still limited in several ways, e.g., it does not yet incorporate document relevance, other viewpoint diversity notions, or the personal preferences and beliefs of users. We encourage researchers and practitioners to build on our work to help improve the measurement of viewpoint diversity in search results. Finally, annotating viewpoints is a difficult, time-consuming task even for expert annotators [15, 20]. Recent work has already applied automatic stance detection methods to search results [22] but did so far not attempt to identify logics of evaluation. However, once such automatic systems have become more comprehensive, researchers and practitioners could easily combine them with existing methods for extracting arguments [12, 60] and visualize viewpoints [4, 17] in search results.

6 Conclusion

We proposed a metric for evaluating viewpoint diversity in search results, measuring the divergence from an ideal scenario of equal viewpoint representation. In a case study evaluating search results on three different debated topics from two

popular search engines, we found that search results may often not be viewpoint-diverse, even if queries are formulated neutrally. We also saw notable differences between search engines concerning bias *magnitude* and *direction*. Our hierarchical viewpoint diversification method, based on HxQuAD, consistently improved the viewpoint diversity of search results. In sum, our results suggest that, while viewpoint bias in search results is not pervasive, users may unknowingly be exposed to high levels of viewpoint bias, depending on the query, topic, or search engine. These factors may influence (especially vulnerable and undecided) users' attitudes by means of recently demonstrated search engine manipulation effects and thereby affect individuals, businesses, and society.

Acknowledgements. This activity is financed by IBM and the Allowance for Top Consortia for Knowledge and Innovation (TKI's) of the Dutch ministry of economic affairs.

References

1. Abid, A., et al.: A survey on search results diversification techniques. *Neural Comput. Appl.* **27**(5), 1207–1229 (2015). <https://doi.org/10.1007/s00521-015-1945-5>
2. Agrawal, R., Gollapudi, S., Halverson, A., Ieong, S.: Diversifying search results. In: *Proceedings of the Second ACM International Conference on Web Search and Data Mining - WSDM 2009*, p. 5. ACM Press, Barcelona, Spain (2009). <https://doi.org/10.1145/1498759.1498766>, <http://portal.acm.org/citation.cfm?doid=1498759.1498766>
3. Ajjour, Y., Alshomary, M., Wachsmuth, H., Stein, B.: Modeling frames in argumentation. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2922–2932. Association for Computational Linguistics, Hong Kong, China (Nov 2019). <https://doi.org/10.18653/v1/D19-1290>, <https://aclanthology.org/D19-1290>
4. Ajjour, Y., et al.: Visualization of the topic space of argument search results in args. me. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 60–65 (2018)
5. Allam, A., Schulz, P.J., Nakamoto, K.: The impact of search engine selection and sorting criteria on vaccination beliefs and attitudes: two experiments manipulating google output. *J. Med. Internet Res.* **16**(4), e100 (Apr 2014). <https://doi.org/10.2196/jmir.2642>, <http://www.jmir.org/2014/4/e100/>
6. Azzopardi, L.: Cognitive Biases in Search: a review and reflection of cognitive biases in information retrieval. In: *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, pp. 27–37. ACM, Canberra ACT Australia (Mar 2021). <https://doi.org/10.1145/3406522.3446023>, <https://dl.acm.org/doi/10.1145/3406522.3446023>
7. Baden, C., Springer, N.: Com(ple)menting the news on the financial crisis: the contribution of news users' commentary to the diversity of viewpoints in the public debate. *Euro. J. Commun.* **29**(5), 529–548 (Oct 2014). <https://doi.org/10.1177/0267323114538724>, <http://journals.sagepub.com/doi/10.1177/0267323114538724>
8. Baden, C., Springer, N.: Conceptualizing viewpoint diversity in news discourse. *Journalism* **18**(2), 176–194 (Feb 2017). <https://doi.org/10.1177/1464884915605028>, <http://journals.sagepub.com/doi/10.1177/1464884915605028>

9. Biega, A.J., Gummadi, K.P., Weikum, G.: Equity of Attention: amortizing individual fairness in rankings. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, pp. 405–414. ACM, Ann Arbor MI USA (Jun 2018). <https://doi.org/10.1145/3209978.3210063>, <https://dl.acm.org/doi/10.1145/3209978.3210063>
10. Bink, M., Zimmerman, S., Elsweiler, D.: Featured snippets and their influence on users' credibility judgements. In: ACM SIGIR Conference on Human Information Interaction and Retrieval, pp. 113–122. CHIIR '22, Association for Computing Machinery, New York, NY, USA (2022). <https://doi.org/10.1145/3498366.3505766>, <https://doi.org/10.1145/3498366.3505766>
11. Boltanski, L., Thévenot, L.: On justification: economies of worth, vol. 27. Princeton University Press (2006)
12. Bondarenko, A., Ajjour, Y., Dittmar, V., Homann, N., Braslavski, P., Hagen, M.: Towards understanding and answering comparative questions. In: Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, pp. 66–74 (2022)
13. Bondarenko, A., et al.: Overview of touché 2021: argument retrieval. In: International Conference of the Cross-Language Evaluation Forum for European Languages, pp. 450–467. Springer (2021)
14. Boykoff, M.T., Boykoff, J.M.: Balance as bias: global warming and the us prestige press. *Glob. Environ. Chang.* **14**(2), 125–136 (2004)
15. Burscher, B., Odijk, D., Vliegenthart, R., De Rijke, M., De Vreese, C.H.: Teaching the computer to code frames in news: comparing two supervised machine learning approaches to frame analysis. *Commun. Methods Measures* **8**(3), 190–206 (2014)
16. Carroll, N.: In Search We Trust: exploring how search engines are shaping society. *Int. J. Knowl. Soc. Res.* **5**(1), 12–27 (Jan 2014). <https://doi.org/10.4018/ijksr.2014010102>, <http://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/ijksr.2014010102>
17. Chamberlain, J., Kruschwitz, U., Hoeber, O.: Scalable visualisation of sentiment and stance. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan (May 2018), <https://aclanthology.org/L18-1660>
18. Chen, S., Khashabi, D., Yin, W., Callison-Burch, C., Roth, D.: Seeing things from a different angle: discovering diverse perspectives about claims. In: Proceedings of NAACL-HLT, pp. 542–557 (2019)
19. Clarke, C.L., et al.: Novelty and diversity in information retrieval evaluation. In: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR 2008, p. 659. ACM Press, Singapore, Singapore (2008). <https://doi.org/10.1145/1390334.1390446>, <http://portal.acm.org/citation.cfm?doid=1390334.1390446>
20. Draws, T., Inel, O., Tintarev, N., Baden, C., Timmermans, B.: Comprehensive viewpoint representations for a deeper understanding of user interactions with debated topics. In: ACM SIGIR Conference on Human Information Interaction and Retrieval, pp. 135–145 (2022)
21. Draws, T., Liu, J., Tintarev, N.: Helping users discover perspectives: enhancing opinion mining with joint topic models. In: 2020 International Conference on Data Mining Workshops (ICDMW), pp. 23–30. IEEE, Sorrento, Italy (Nov 2020). <https://doi.org/10.1109/ICDMW51313.2020.00013>, <https://ieeexplore.ieee.org/document/9346407/>

22. Draws, T., et al.: Explainable cross-topic stance detection for search results. In: CHIIR 2023: ACM SIGIR Conference on Human Information Interaction and Retrieval. CHIIR 2023, ACM SIGIR Conference on Human Information Interaction and Retrieval (2023)
23. Draws, T., Tintarev, N., Gadiraju, U., Bozzon, A., Timmermans, B.: Assessing viewpoint diversity in search results using ranking fairness metrics. *ACM SIGKDD Explorations Newsletter* 23(1), 50–58 (May 2021). <https://doi.org/10.1145/3468507.3468515>, <https://dl.acm.org/doi/10.1145/3468507.3468515>
24. Draws, T., Tintarev, N., Gadiraju, U., Bozzon, A., Timmermans, B.: This is not what we ordered: exploring why biased search result rankings affect user attitudes on debated topics. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 295–305. ACM, Virtual Event Canada (Jul 2021). <https://doi.org/10.1145/3404835.3462851>, <https://dl.acm.org/doi/10.1145/3404835.3462851>
25. Drosou, M., Pitoura, E.: Search result diversification. *SIGMOD Record* 39(1), 7 (2010)
26. Dumani, L., Neumann, P.J., Schenkel, R.: A framework for argument retrieval. In: Jose, J.M., et al. (eds.) *ECIR 2020*. LNCS, vol. 12035, pp. 431–445. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-45439-5_29
27. Epstein, R., Robertson, R.E.: The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. In: Proceedings of the National Academy of Sciences 112(33), E4512–E4521 (Aug 2015). <https://doi.org/10.1073/pnas.1419828112>, <http://www.pnas.org/lookup/doi/10.1073/pnas.1419828112>
28. Epstein, R., Robertson, R.E., Lazer, D., Wilson, C.: Suppressing the search engine manipulation effect (SEME). In: Proceedings of the ACM on Human-Computer Interaction 1(CSCW), 1–22 (Dec 2017). <https://doi.org/10.1145/3134677>, <https://dl.acm.org/doi/10.1145/3134677>
29. Fuglede, B., Topsoe, F.: Jensen-shannon divergence and hilbert space embedding. In: International Symposium on Information Theory, 2004. ISIT 2004. Proceedings, p. 31. IEEE (2004)
30. Gao, R., Shah, C.: Toward creating a fairer ranking in search engine results. *Inf. Process. Manag.* 57(1), 102138 (Jan 2020). <https://doi.org/10.1016/j.ipm.2019.102138>, <https://linkinghub.elsevier.com/retrieve/pii/S0306457319304121>
31. Gevelber, L.: It’s all about ‘me’-how people are taking search personally. Tech. rep. (2018). <https://www.thinkwithgoogle.com/marketing-strategies/search/personal-needs-search-trends/>
32. Gezici, G., Lipani, A., Saygin, Y., Yilmaz, E.: Evaluation metrics for measuring bias in search engine results. *Inf. Retrieval J.* 24(2), 85–113 (Apr 2021). <https://doi.org/10.1007/s10791-020-09386-w>, <http://link.springer.com/10.1007/s10791-020-09386-w>
33. Ghenai, A., Smucker, M.D., Clarke, C.L.: A think-aloud study to understand factors affecting online health search. In: Proceedings of the 2020 Conference on Human Information Interaction and Retrieval, pp. 273–282. ACM, Vancouver BC Canada (Mar 2020). <https://doi.org/10.1145/3343413.3377961>, <https://dl.acm.org/doi/10.1145/3343413.3377961>
34. Grady, C., Lease, M.: Crowdsourcing document relevance assessment with mechanical turk. In: NAACL HLT Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk, pp. 172–179 (2010)
35. Gretz, S., et al.: A large-scale dataset for argument quality ranking: construction and analysis. *Proc. AAAI Conf. Artif. Intell.* 34, 7805–7813 (2020). <https://doi.org/10.1609/aaai.v34i05.6285>

36. Han, B., Shah, C., Saelid, D.: Users' perception of search-engine biases and satisfaction. In: Boratto, L., Faralli, S., Marras, M., Stilo, G. (eds.) *BIAS 2021*. CCIS, vol. 1418, pp. 14–24. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-78818-6_3
37. Helberger, N.: On the democratic role of news recommenders. *Digital Journalism* 7(8), 993–1012 (Sep 2019). <https://doi.org/10.1080/21670811.2019.1623700>, <https://www.tandfonline.com/doi/full/10.1080/21670811.2019.1623700>
38. Hu, S., Dou, Z., Wang, X., Sakai, T., Wen, J.R.: Search result diversification based on hierarchical intents. In: *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pp. 63–72. ACM, Melbourne Australia (Oct 2015). <https://doi.org/10.1145/2806416.2806455>, <https://dl.acm.org/doi/10.1145/2806416.2806455>
39. Joachims, T., Granka, L., Pan, B., Hembrooke, H., Gay, G.: Accurately interpreting clickthrough data as implicit feedback. *ACM SIGIR Forum* 51(1), 8 (2016)
40. Kaya, M., Bridge, D.: Subprofile-aware diversification of recommendations. *User Modeling and User-Adapted Interaction* 29(3), 661–700 (Jul 2019). <https://doi.org/10.1007/s11257-019-09235-6>, <http://link.springer.com/10.1007/s11257-019-09235-6>
41. Küçük, D., Can, F.: Stance detection: a survey. *ACM Comput. Surv. (CSUR)* 53(1), 1–37 (2020)
42. Kulshrestha, J., et al.: Quantifying search bias: investigating sources of bias for political searches in social media. In: *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pp. 417–432. ACM, Portland Oregon USA (Feb 2017). <https://doi.org/10.1145/2998181.2998321>, <https://dl.acm.org/doi/10.1145/2998181.2998321>
43. Kulshrestha, J., et al.: Search bias quantification: investigating political bias in social media and web search. *Information Retrieval Journal* 22(1–2), 188–227 (Apr 2019). <https://doi.org/10.1007/s10791-018-9341-2>, <http://link.springer.com/10.1007/s10791-018-9341-2>
44. Ludolph, R., Allam, A., Schulz, P.J.: Manipulating google's knowledge graph box to counter biased information processing during an online search on vaccination: application of a technological debiasing strategy. *J. Med. Internet Res.* 18(6), e137 (Jun 2016). <https://doi.org/10.2196/jmir.5430>, <http://www.jmir.org/2016/6/e137/>
45. McKay, D., et al.: We are the change that we seek: information interactions during a change of viewpoint. In: *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*, pp. 173–182 (2020)
46. McKay, D., Owyong, K., Makri, S., Gutierrez Lopez, M.: Turn and face the strange: investigating filter bubble bursting information interactions. In: *ACM SIGIR Conference on Human Information Interaction and Retrieval*, pp. 233–242. CHIIR '22, Association for Computing Machinery, New York, NY, USA (2022). <https://doi.org/10.1145/3498366.3505822>
47. Mulder, M., Inel, O., Oosterman, J., Tintarev, N.: Operationalizing framing to support multiperspective recommendations of opinion pieces. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 478–488. ACM, Virtual Event Canada (Mar 2021). <https://doi.org/10.1145/3442188.3445911>, <https://dl.acm.org/doi/10.1145/3442188.3445911>
48. Munson, S.A., Resnick, P.: Presenting diverse political opinions: how and how much. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1457–1466. CHI 2010, Association for Computing Machinery, New York, NY, USA (2010). <https://doi.org/10.1145/1753326.1753543>

49. Pan, B., Hembrooke, H., Joachims, T., Lorigo, L., Gay, G., Granka, L.. In *Google We Trust: Users' Decisions on Rank, Position, and Relevance*. *J. Comput.-Mediated Commun.* 12(3), 801–823 (Apr 2007). <https://doi.org/10.1111/j.1083-6101.2007.00351.x>, <https://academic.oup.com/jcmc/article/12/3/801-823/4582975>
50. Pathiyar Cherumanal, S., Spina, D., Scholer, F., Croft, W.B.: Evaluating fairness in argument retrieval. In: *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pp. 3363–3367 (2021)
51. Pennycook, G., Rand, D.G.: Lazy, not biased: susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition* 188, 39–50 (Jul 2019). <https://doi.org/10.1016/j.cognition.2018.06.011>, <https://linkinghub.elsevier.com/retrieve/pii/S001002771830163X>
52. Pogacar, F.A., Ghenai, A., Smucker, M.D., Clarke, C.L.: The positive and negative influence of search results on people's decisions about the efficacy of medical treatments. In: *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*, pp. 209–216. ACM, Amsterdam The Netherlands (Oct 2017). <https://doi.org/10.1145/3121050.3121074>, <https://dl.acm.org/doi/10.1145/3121050.3121074>
53. Purcell, K., Rainie, L., Brenner, J.: *Search engine use 2012* (2012)
54. Puschmann, C.: Beyond the bubble: assessing the diversity of political search results. *Digital Journalism* 7(6), 824–843 (Jul 2019). <https://doi.org/10.1080/21670811.2018.1539626>, <https://www.tandfonline.com/doi/full/10.1080/21670811.2018.1539626>
55. Radlinski, F., Craswell, N.: Comparing the sensitivity of information retrieval metrics. In: *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 667–674 (2010)
56. Reimer, J.H., Huck, J., Bondarenko, A.: Grimjack at touché 2022: axiomatic re-ranking and query reformulation. *Working Notes Papers of the CLEF* (2022)
57. Rieger, A., Draws, T., Theune, M., Tintarev, N.: This item might reinforce your opinion: obfuscation and labeling of search results to mitigate confirmation bias. In: *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*, pp. 189–199 (2021)
58. Sakai, T., Craswell, N., Song, R., Robertson, S., Dou, Z., Lin, C.Y.: Simple evaluation metrics for diversified search results, p. 9 (2010)
59. Santos, R.L., Macdonald, C., Ounis, I.: Exploiting query reformulations for web search result diversification. In: *Proceedings of the 19th international conference on World wide web*, pp. 881–890 (2010)
60. Stab, C., et al.: ArgumenText: searching for arguments in heterogeneous sources. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pp. 21–25 (2018)
61. Tintarev, N., Sullivan, E., Guldin, D., Qiu, S., Odjik, D.: Same, same, but different: algorithmic diversification of viewpoints in news. In: *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization*, pp. 7–13. ACM, Singapore Singapore (Jul 2018). <https://doi.org/10.1145/3213586.3226203>, <https://dl.acm.org/doi/10.1145/3213586.3226203>
62. Voorhees, E.M.: Variations in relevance judgments and the measurement of retrieval effectiveness. *Inf. Process. Manag.* **36**(5), 697–716 (2000)
63. Vrijenhoek, S., Bénédicte, G., Gutierrez Granada, M., Odijk, D., De Rijke, M.: Radio-rank-aware divergence metrics to measure normative diversity in news recommendations. In: *Proceedings of the 16th ACM Conference on Recommender Systems*. pp. 208–219 (2022)

64. Vrijenhoek, S., Kaya, M., Metoui, N., Möller, J., Odijk, D., Helberger, N.: Recommenders with a mission: Assessing diversity in news recommendations. In: Proceedings of the 2021 Conference on Human Information Interaction and Retrieval, pp. 173–183. CHIIR 2021, Association for Computing Machinery, New York, NY, USA (2021). <https://doi.org/10.1145/3406522.3446019>
65. White, R.: Beliefs and biases in web search. In: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 3–12. ACM, Dublin Ireland (Jul 2013). <https://doi.org/10.1145/2484028.2484053>, <https://dl.acm.org/doi/10.1145/2484028.2484053>
66. White, R.W., Hassan, A.: Content bias in online health search. *ACM Transactions on the Web* 8(4), 1–33 (Nov 2014). <https://doi.org/10.1145/2663355>, <https://dl.acm.org/doi/10.1145/2663355>
67. White, R.W., Horvitz, E.: Belief dynamics and biases in web search. *ACM Trans. Inf. Syst.* 33(4), 1–46 (May 2015). <https://doi.org/10.1145/2746229>, <https://dl.acm.org/doi/10.1145/2746229>
68. Xu, L., Zhuang, M., Gadiraju, U.: How do user opinions influence their interaction with web search results?, pp. 240–244. Association for Computing Machinery, New York, NY, USA (2021), <https://doi.org/10.1145/3450613.3456824>
69. Yamamoto, Y., Shimada, S.: Can disputed topic suggestion enhance user consideration of information credibility in web search? In: Proceedings of the 27th ACM Conference on Hypertext and Social Media, pp. 169–177. ACM, Halifax Nova Scotia Canada (Jul 2016). <https://doi.org/10.1145/2914586.2914592>, <https://dl.acm.org/doi/10.1145/2914586.2914592>
70. Yang, K., Stoyanovich, J.: Measuring fairness in ranked outputs. In: Proceedings of the 29th International Conference on Scientific and Statistical Database Management, pp. 1–6. ACM, Chicago IL USA (Jun 2017). <https://doi.org/10.1145/3085504.3085526>, <https://dl.acm.org/doi/10.1145/3085504.3085526>
71. Yom-Tov, E., Dumais, S., Guo, Q.: Promoting civil discourse through search engine diversity. *Soc. Sci. Comput. Rev.* 32(2), 145–154 (Apr 2014). <https://doi.org/10.1177/0894439313506838>, <http://journals.sagepub.com/doi/10.1177/0894439313506838>
72. Zehlike, M., Bonchi, F., Castillo, C., Hajian, S., Megahed, M., Baeza-Yates, R.: FA*IR: A Fair Top-k ranking algorithm. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, pp. 1569–1578. ACM, Singapore Singapore (Nov 2017). <https://doi.org/10.1145/3132847.3132938>, <https://dl.acm.org/doi/10.1145/3132847.3132938>
73. Zehlike, M., Yang, K., Stoyanovich, J.: Fairness in ranking, part i: score-based ranking. *ACM Comput. Surv.* 55(6), 1–36 (2022)