Explaining Search Result Stances to Opinionated People

Explaining Search Result Stances to Opinionated People

Zhangyi Wu[1](✉), Tim Draws[2], Federico Cau[1], Francesco Barile[1],
Alisa Rieger[2], and Nava Tintarev[1]

[1] Maastricht University, Maastricht, Netherlands
zhangyi.wu@student.maastrichtuniversity.nl,
{federico.cau,f.barile,n.tintarev}@maastrichtuniversity.nl
[2] Delft University of Technology, Delft, Netherlands
{t.a.draws,a.rieger}@tudelft.nl

Abstract. People use web search engines to find information before forming opinions, which can lead to practical decisions with different levels of impact. The cognitive effort of search can leave opinionated users vulnerable to cognitive biases, e.g., the *confirmation bias*. In this paper, we investigate whether stance labels and their explanations can help users consume more diverse search results. We automatically classify and label search results on three topics (i.e., *intellectual property rights*, *school uniforms*, and *atheism*) as *against*, *neutral*, and *in favor*, and generate explanations for these labels. In a user study (N=203), we then investigate whether search result stance bias (balanced vs biased) and the level of explanation (plain text, label only, label and explanation) influence the diversity of search results clicked. We find that stance labels and explanations lead to a more diverse search result consumption. However, we do not find evidence for systematic opinion change among users in this context. We believe these results can help designers of search engines to make more informed design decisions.

Keywords: Explainable Search · Confirmation Bias · User Study

1 Introduction

Web search that can lead to consequential decision-making frequently concerns *debated topics*, topics that different people and groups disagree on, such as *whether to vaccinate a child* or *whether nuclear energy should be used as a power source.* Prior research has shown that the interplay between search engine biases and users' cognitive biases can lead to noteworthy behavioral patterns. For instance, when search result stance biases interact with cognitive user biases, information seekers may experience the *search engine manipulation effect* (SEME): the tendency to adopt the stance expressed by the majority of (highly-ranked) search results [3,9,15,33]. However, these results have only been studied and found for users who are undecided, not for users who already have strong opinions, who we refer to as *opinionated* users.

L. Longo (Ed.): xAI 2023, CCIS 1902, pp. 573–596, 2023.
https://doi.org/10.1007/978-3-031-44067-0_29

High cognitive demand during complex web searches can increase the risk of cognitive biases [6]. One such bias is the *confirmation bias*, which involves a preference for information that aligns with preexisting beliefs while disregarding contradictory information during the search process [6,31]. Interventions to mitigate confirmation bias during web search have aimed at decreasing engagement with attitude-confirming and increasing engagement with attitude-opposing information [38], i.e., reducing interaction with search results that confirm a user's attitude and increasing interaction with search results that challenge a user's attitude. Interventions to reduce interaction with particular items have also been investigated in the context of misinformation. One effective method for reducing interaction with misleading content involves *labels* to flag certain items [10,23,29].

The core issue with confirmation bias during web search, similar to the related issues of misinformation and SEME, is that users consume biased content. This motivated us to investigate interventions to increase the diversity of consumption and specifically whether labels indicating stance, and their explanations, are likewise successful for confirmation bias mitigation during search on debated topics. Therefore the goal of these interventions is to promote unbiased web search and mitigate the effects of users' confirmation bias and underlying (stance) biases in a *search engine result page* (SERP). Consequently, this paper aims to address the following question: ***Do automatically generated stance labels and explanations of the labels for search results increase the diversity of viewpoints users engage with, even if search results are biased?***

To address this question, we test three hypotheses. Previous work has found that undecided users are likely to change their opinion when exposed to biased search results, since they select more search results reflecting a certain opinion [3,14,15,33]. However, in this study we restrict participants to opinionated users, having *strong* existing opinions on a topic, and investigate whether ***H1a):*** *Users who are exposed to viewpoint-biased search results interact with* <u>*less diverse results*</u> *than users who are exposed to balanced search results.*

Second, informative labels have been shown to mitigate confirmation bias in search results [38]. Therefore, in this study, we investigate whether simple stance labels (*against*, *neutral*, and *in favor*), and stance labels with explanations (importance of keywords) are effective for mitigating confirmation bias. This leads to ***H1b)****: Users who are exposed to search results with (1) stance labels or (2) stance labels with explanations for each search result interact with more diverse content than users who are exposed to regular search results.*

Third, if the labels are effective in reducing confirmation bias, we would expect an interaction effect between the bias in search results and explanation level (plain, label, label plus explanation): ***H1c)*** *Users who are exposed to search results with (1) stance labels or (2) stance labels with explanations are less susceptible to the effect of viewpoint biases in search results on clicking diversity.*

We investigate these hypotheses in a pre-registered between-subjects user study (N=203) simulating an open search task.[1] Our results show that both

[1] The pre-registration is openly available at https://osf.io/3nxak.

stance labels and explanations, led to a more diverse search result consumption compared to plain (unlabeled) search result pages. However, we did not find evidence that the explanations influenced opinion change. We believe these results can help designers of search engines to make more informed design decisions.

2 Related Work

Explainable Artificial Intelligence (XAI) aims to help people understand the decisions and predictions AI systems make. In this paper, we investigate specifically how XAI can support users in searching for disputed topics. Search for debated topics is highly subjective: when users search the web to seek advice or form opinions on these kinds of topics, not just search result *relevance* but also the stance of content is influential [3,14,15,33,38]. To mitigate undesired effects such as biased opinion change, earlier work has measured and increased the fairness [18,50,53] and viewpoint diversity in search results [13,34,47].

On the user interface side, it could be fruitful to label and explain the stance represented on a search engine results page (or SERP). These labels are related to the task known as *stance detection*, which is predominantly applied in a *target-specific* fashion. That is, detecting not just a sentiment, but how it is referred to in relation to a specific topic or claim (often referred to as the *target*, e.g., "people should wear school uniforms") [2]. Stance detection is a multi-class classification task (i.e., typically classifying documents into *against*, *neutral*, and *in favor*, so predictive performances are most commonly reported in terms of macro F1 scores [24]. Furthermore, web search interventions targeting the mitigation of undesired effects, such as SEME, require *cross-target* stance detection models to quickly respond to the large variety of debated topics users may search for. Here, stance detection models are applied to data sets where each document may be relevant to one of many potential topics [2,24]. Constructing models that classify documents into stances related to *any* topic in such a way may lead to weaker predictive accuracy compared to target-specific methods, but makes stance detection more generalizable and scalable. Cross-target ternary stance detection by previous work (e.g., on news articles or tweets) have ranged roughly from macro F1 scores of .450 to .750 [1,4,5,21,36,49].

Also comparable are the cross-topic stance detection models evaluated using the *Emergent* data set (and its follow-up version, the *2017 Fake News Challenge* data set) which have achieved macro F1 scores of up to .756 [20,39,41]. While the main contribution of this paper is not to improve on the state of the art for stance detection, the stance detection method (DistilBERT) used here is comparable to this state of the art (macro F1 of 0.72). DistilBERT is much smaller than other pre-trained models, and handles small datasets well [42,44,46].

What XAI methods are suitable for explaining stance detection to users? Stance detection can be seen as a text classification task. For text classification, explanations containing *input features* have been found to be highly adaptable and often meaningful to humans [12,28].

The way in which explanations can be visualized depends on the data type, purpose, and audience. Many current methods indicate input features as feature importance using a saliency map [37,45]. When the features are readable texts, saliency information is shown by highlighting the most significant input at word or token level [22]. There have been some instances where researchers used text-based saliency maps to demonstrate their findings [19]. To the best of our knowledge, no previous work has explored whether highlighting salient words in search results would mitigate people's clicking bias. One of the most closely related works found that feature-based explanations could help users simulate model predictions for search results [12]. Another similar study involves a verbal saliency map using a model-agnostic explainer and a human evaluation of explanation representations of news topic classifier and sentiment analysis [17]. Their finding is that the saliency map makes explanations more understandable and less cognitively challenging for humans than heatmap visualization. However, our work differs from theirs in several ways: we study explanations in the context of search engines, and we have conducted a full-fledged user study while they only performed a pilot study.

Contribution to Knowledge for XAI. Previous XAI literature has contributed to explaining information retrieval systems, focusing on the interpretability of document-retrieval mechanisms [25,26,52]. For example, the authors of [52] propose a listwise explanation generator, which provides an explanation that covers all the documents contained in the page (e.g., by describing which query aspects were covered by each document). These explanations were not evaluated by people. Another paper studied how well explanations of individual search results helped people anticipate model predictions [12], but did not consider cognitive or stance bias. In contrast to previous work, this paper examines how explanations affect users, considering the potential mitigation of their cognitive biases on search engine manipulation effects (SEME). In doing so, we see a need to address both potential (stance) bias within a search result page, and the bias of users consuming these results. To the best of our knowledge, this is also the first work to conduct empirical experiments on users' *behavior* in response to explanations in the context of information retrieval.

3 Methodology

This section describes the materials we used for organizing the user study (e.g., data set, stance detection model, explanation generation, and search interface).

3.1 Data Preparation

To train, test, and explain the stance detection model, we considered a public data set containing search results related to three debated topics (i.e., *atheism*, *intellectual property rights*, and *school uniforms*) [13].[2] [13] motivate the selection

[2] The data set is available at https://osf.io/yghr2.

Table 1. Topic and stance distribution in the used data.

Topic	N	Stance Distribution Against – Neutral – In Favor
Intellectual property rights	378	10.5% – 17.7% – 71.7%
School uniforms	395	21.5% – 36.7% – 41.8%
Atheism	352	19.8% – 46.3% – 33.8%
Total	1125	17.3% – 33.3% – 49.3%

of these three topics because they offer valid arguments for both supporting and opposing viewpoints. Additionally, they argue that opinions on these topics have diverse impacts, ranging from concerning mainly the user (atheism) to businesses (intellectual property rights) and society (school uniforms). These data include URLs, titles, snippets, and stance labels for a total of 1475 search results, which had been retrieved via API or web crawling from two popular web search engines.

Stance labels had been assigned (by experts) on seven-point Likert scales (i.e., including three degrees of opposing or supporting a topic), which we mapped into the three categories *against*, *neutral*, and *in favor* (i.e., which is what most current stance detection methods handle). Using the provided URLs, we crawled the full web page text bodies (stripped of any HTML tags) for all search results. We here dropped 347 search results from the data as their text bodies could not be retrieved (e.g., because of 404 errors), leaving 1125 search results accompanied by their respective text bodies. Finally, we processed the retrieved contents by truncating each document's middle section, retaining only its head and tail, then concatenating each search result's title, snippet, and the head section and tail section, while ensuring that the result is exactly 510 tokens long. We removed all other information from the data aside from the documents' stance labels. Table 1 shows the stance distribution per topic in our final data set.

3.2 Stance Detection Model

After pre-processing (tokenization), we developed the model for classifying search results into against, neutral, and in favor. The dataset was split into training (75%) and validation/test (25%) sets. We fine-tuned the model for different hyperparameters. After every epoch the validation set were evaluated to monitor its learning progress. Once the model's evaluation loss stops to decrease for 5 epochs the training is terminated and is evaluated based on the unseen test set, and the result is considered the general performance that we report. The best-performing model and learned parameters were used as the predictor for identifying labels and generating explanations.

Tokenization. Before training the stance classification model, we needed further preprocessing of the raw search results. Specifically, we had to *tokenize* each word before feeding the search results into the model. The tokenization

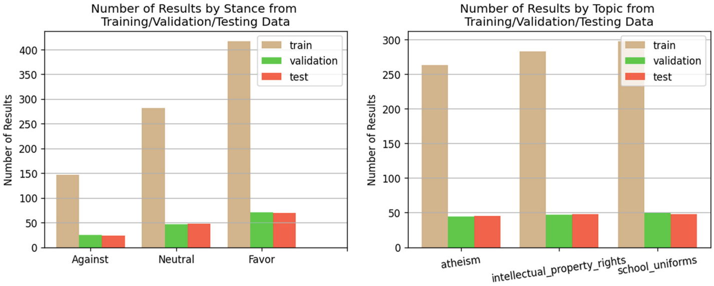


Fig. 1. Distributions of Attributes Across the Training, Validation and Test Set

process performs tasks such as handling subwords, adding special tokens to pad a sequence of words to a `max length`, or indicating the beginning or end of a sequence. In our work, we instantiated the DistilBERT tokenizer using the AutoTokenizer class provided by the `transformers` [48]. We set the `length` of the tokenizer to a size of 512.

Training Details. Similar to previous work performing search result stance classification [12], we built one cross-topic model using our entire data set. First, we split the dataset in a stratified manner, which preserves the percentage of the topic and labels in each subset. Figure 1 shows the final split across the training, validation, and test sets on both topics and labels. Then, we classified search results into stance categories (i.e., *in favor*, *against*, or *neutral*), using a pre-trained version of the uncased DistilBERT base model [40] from HuggingFace. Specifically, we fine-tuned DistilBERT using 75% (843) of the documents in our data and split the remaining 25% (282) equally for validation and testing. Due to the relatively small size of our dataset, we tried to avoid over-fitting by using neural network dropout and stopping training when evaluation loss stops improving for 5 epochs [43,51]. We trained using the same dataset split and experimented with learning rates ranging from 5*e*–6 to 1*e*–4. Regarding the remain hyper-parameters, the models were optimized using the Adam optimizer with a learning rate of $1e-5$, the batch size was set to 8 and we set the dropout rate of both attention layers and fully connected layers to 0.4.

Metric. In our stance detection task, we have three labels to be classified, and their distribution is uneven. To take performance on all stances into account, we considered the macro F1 score, defined as the average of the F1 scores for each of the three stances:

$$macro_{F1} = (F1(stance = favor) + F1(stance = neutral) + F1(stance = against))/3 \quad (1)$$

Model Performance. We obtained a stance detection model by fine-tuning the DistilBERT model for our downstream classification task. The fine-tuning process was completed through HuggingFace's `Trainer` interface. The progress of each run was tracked using Weights & Biases.[3] We observed that the learning rate of 1e–5 gave the best performance with a macro F1 score of 0.72.[4]

3.3 Creating Explanations

Instance Selection. For the user study, we selected search results correctly predicted by the model, picking them mainly from the test set and some from the validation set. As mentioned before, the train, validation, and test sets were split in a stratified way, which preserves the frequencies of our target classes. To assemble search results for our study, we randomly drew 21 *correctly predicted* search results per topic from our test and validation data (i.e., seven against, seven neutral toward, and seven in favor of the topic). The SERPs later displayed 10 results per page.

LIME Parameter Tuning. After we fine-tuned and evaluated the model on the search result corpus, we used LIME (Local Interpretable Model-Agnostic Explanations) to explain the model's predictions [37].[5] For text data, LIME outputs a list of features (tokens) from the input sentence, ranked by their importance w.r.t. a model's specific prediction. We generated the explanations by setting the neighborhood size to learn the linear model to 5000, kernel_width to 50,[6] and showing the top 20 important tokens (or less based on the text length) belonging to the model's predicted class.

3.4 Search Engine

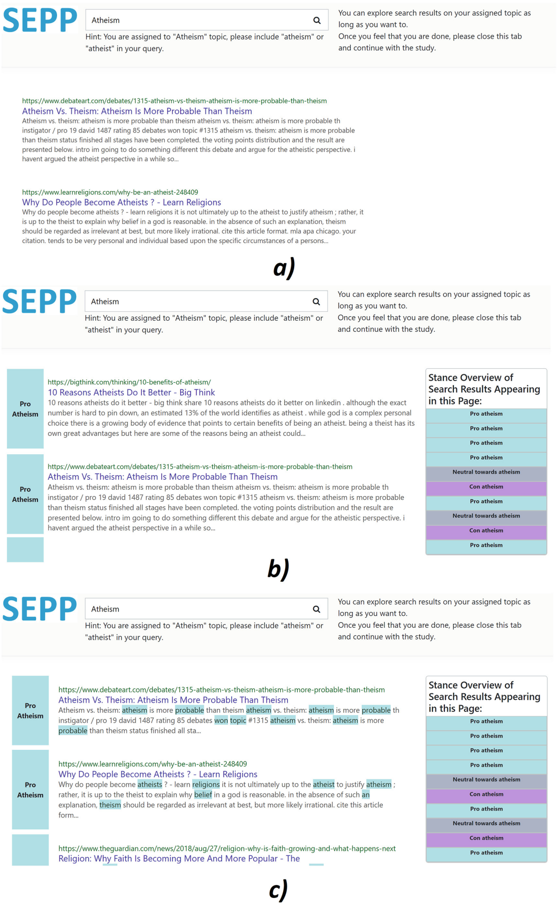
Architecture. We implemented our web-based search interface using the SearchX platform as a basis [35]. The web application has both a front-end and a back-end. The front-end is built on NodeJS using React and Flux frameworks and manages users' data by sending logs to the back-end, which mainly handles the retrieval of search results from the database and stores the logs from the front-end.

[3] https://wandb.ai.

[4] Due to a minor error in evaluation, a slightly higher macro F1 score was reported in the pre-registration. However, this erroneous score did not influence the training process or affect our user study.

[5] https://github.com/marcotcr/lime.

[6] We tried multiple kernel sizes (10, 25, 50, and 75) and chose a value of 50 since we got a slight increase in the R2 scores for each LIME local prediction on the test set of about 3–4% on average compared to the other sizes.



Hint: You are assigned to "Atheism" topic, please include "atheism" or "atheist" in your query.

You can explore search results on your assigned topic as long as you want to.
Once you feel that you are done, please close this tab and continue with the study.

a)

https://www.debateart.com/debates/1315-atheism-vs-theism-atheism-is-more-probable-than-theism
Atheism Vs. Theism: Atheism Is More Probable Than Theism
Atheism vs. theism: atheism is more probable than theism atheism vs. theism: atheism is more probable th instigator / pro 19 david 1487 rating 85 debates won topic #1315 atheism vs. theism: atheism is more probable than theism status finished all stages have been completed. the voting points distribution and the result are presented below. intro im going to do something different this debate and argue for the atheistic perspective. i havent argued the atheist perspective in a while so...

https://www.learnreligions.com/why-be-an-atheist-248409
Why Do People Become Atheists ? - Learn Religions
Why do people become atheists ? - learn religions it is not ultimately up to the atheist to justify atheism ; rather, it is up to the theist to explain why belief in a god is reasonable. in the absence of such an explanation, theism should be regarded as irrelevant at best, but more likely irrational. cite this article format. mla apa chicago. your citation. tends to be very personal and individual based upon the specific circumstances of a persons...

b)

https://bigthink.com/thinking/10-benefits-of-atheism/
10 Reasons Atheists Do It Better - Big Think
10 reasons atheists do it better - big think share 10 reasons atheists do it better on linkedin . although the exact number is hard to pin down, an estimated 13% of the world identifies as atheist . while god is a complex personal choice there is a growing body of evidence that points to certain benefits of being an atheist. being a theist has its own great advantages but here are some of the reasons being an atheist could...

https://www.debateart.com/debates/1315-atheism-vs-theism-atheism-is-more-probable-than-theism
Atheism Vs. Theism: Atheism Is More Probable Than Theism
Atheism vs. theism: atheism is more probable than theism atheism vs. theism: atheism is more probable th instigator / pro 19 david 1487 rating 85 debates won topic #1315 atheism vs. theism: atheism is more probable than theism status finished all stages have been completed. the voting points distribution and the result are presented below. intro im going to do something different this debate and argue for the atheistic perspective. i havent argued the atheist perspective in a while so...

c)

https://www.debateart.com/debates/1315-atheism-vs-theism-atheism-is-more-probable-than-theism
Atheism Vs. Theism: Atheism Is More Probable Than Theism
Atheism vs. theism: atheism is more probable than theism atheism vs. theism: atheism is more probable th instigator / pro 19 david 1487 rating 85 debates won topic #1315 atheism vs. theism: atheism is more probable than theism status finished all sta...

https://www.learnreligions.com/why-be-an-atheist-248409
Why Do People Become Atheists ? - Learn Religions
Why do people become atheists ? - learn religions it is not ultimately up to the atheist to justify atheism ; rather, it is up to the theist to explain why belief in a god is reasonable. in the absence of such an explanation, theism should be regarded as irrelevant at best, but more likely irrational. cite this article form...

https://www.theguardian.com/news/2018/aug/27/religion-why-is-faith-growing-and-what-happens-next
Religion: Why Faith Is Becoming More And More Popular - The

Fig. 2. Different conditions for the SERP display. *a)* Text-only Search Result Page. *b)* Predicted stance labels. *c)* Predicted stance labels and Explanations.

Interface. We designed the search interface as follows. As soon as users open the homepage of the study, saw a search bar positioned at the top of the page, where they may enter their queries. The search engine showed results only when the input query included one or more keywords referring to the user's assigned topic (among atheism, intellectual property rights, and school uniforms). Otherwise, showed a message informing users that no results were found for their query. We provided users with the topic and the specific keyword to include in the query with a sentence below the search bar. For the *Atheism* topic, the keywords are "atheist" or "atheism", for the *intellectual property right* topic, the keywords are "property right", "intellectual right" or "intellectual property right", while for the *school uniform* topic, keywords are "uniform" or "school uniform".

After the users inserted a query including the above mentioned keywords, the interface displayed a list of matched results. This set of search results had a different arrangement based on the random condition a user was assigned to. We set up three SERP display conditions (see Sect. 4.2): 1) text-only SERP, 2) SERP with labels; and 3) SERP with labels and explanations. The text-only SERP showed ten search results without any extra information. Each search result had two parts: a clickable title redirecting the user to the corresponding webpage and a snippet showing some of its content. Figure 2 a) shows the layout of this type of interface. The SERP with labels introduces two new features: the labels indicating the original content's stance and an overview of all the search results labels on the right side of the webpage. As shown in Fig. 2 b), we assigned a different color representing each viewpoint. We also added extra information to each label, indicating whether the document is *against*, *neutral*, or *in favor* of the topic. The third type of SERP interface makes use of a mono-chromatic saliency map (see Fig. 2 c), highlighting the top 20 words that best contribute to the prediction in the search result snippet. Due to the space limitation of the interface, not all the feature would appear in the snippet. The color of the saliency map aligns with the color of the label. We use the same color for all 20 feature words regardless of their significance to make it less complicated for users to understand.

Table 2. Templates for the first SERP (i.e., the top 10-ranked search results) in each SERP ranking bias condition.

Rank	Biased Opp. (T1)	Biased Supp. (T2)	Balanced (T3)
1	Against	Favor	Neutral
2	Against	Favor	Neutral
3	Against	Favor	Neutral
4	Against	Favor	Neutral
5	Favor	Against	Against
6	Neutral	Neutral	Neutral
7	Against	Favor	Favor
8	Favor	Against	Against
9	Neutral	Neutral	Neutral
10	Against	Favor	Favor

Biased Versus Balanced Results. We ranked search results depending on the ranking bias condition randomly assigned to each user. Specifically, we created *biased* and *balanced* top 10 search result ranking templates, according to which we would later display search results to users. Biased search results were biased either in the *against* or *in favor* direction and thus contain only results of one of these two stance classes in the top four ranking spots (the remaining six results were balanced across stance classes). Users who were assigned to the *biased* condition would see search results biased in the opposite direction compared to their pre-search opinion (e.g., if they were in favor of school uniforms, they would see results biased against school uniforms). Users who were assigned to the balanced condition would see a similar search result page, but with neutral search results in the top four spots. Neutral results either do not contain any arguments on the debated topic or equally many arguments in both directions.[7] Table 2 shows an overview of the ranking templates.

4 User Study Setup

4.1 Research Ethics and Pre-registration

We deployed the web application on servers owned by the Faculty of Maastricht University (UM server) and secured the connection through HTTPS protocol with SSL certificates. The study was reviewed and accepted by our review board. The data collected from users, ranging from demographical information to user's clicking behaviours, are all stored anonymously in the server. Prior to the launch of the user study, the research question, hypotheses, methodology, measurements, etc. were pre-registered on the Open Science Framework. Only minor changes were made, including balancing the number of participants in the different conditions, change the way we measure users' attitude change, and correcting a computation error in the macro F1 score.

4.2 Variables

In our study, each subject could look at search results (i.e., 10 search results per page) accompanied by different features. We analyzed the participants' attitudes and interaction behavior (with a focus on the proportion of clicks on attitude-confirming search results).

Independent variables.

- *Topic* (between-subjects, categorical). Participants were assigned to one topic (i.e., atheism, intellectual property rights, or school uniforms) for which they have a strong pre-search attitude (i.e., strongly opposing or strongly supporting). If a participant had no strong attitude on any topic, they ended the study. If a participant had multiple strong attitudes, they were assigned to the topic that has the fewest participants at that point in the study (i.e., to move toward a balanced topic distribution).

[7] We chose this setup to make the conditions as comparable as possible, e.g., rather than displaying results in alternating fashion in the balanced condition.

- *SERP ranking bias* (between-subjects, categorical). There were two types of ranking conditions: biased and balanced. For each of these two conditions, we preset a ranking template (see Table 2). Participants would see a search result page with ten items which were ranked in accordance with the template. If a user was assigned to the biased condition, they would see opposing-biased search results if their pre-search attitude was *in favor* (i.e., 3) and supporting-biased search results if their pre-search attitude was *against* (i.e., -3).
- *SERP display* (between-subjects, categorical). Each participant saw search results accompanied by one of these features: (1) plain text results without stance labels, (2) results with predicted stance labels (Fig. 2 b), or (3) results with predicted stance labels and highlighted explanations (Fig. 2 c).

Dependent Variable.

- *Shannon Index* (numerical). The Shannon Index was applied to measure the diversity of users' clicks. Let N be the total number of clicks made in one session, n_0, n_1, n_2 be the number of clicks of "against", "neutral" and "favor" items respectively. The formula for computing clicking diversity is: $-\sum_{i=0}^{2} \frac{n_i}{N} \ln(\frac{n_i}{N})$. The convention for no occurrence of a class is to ignore it as $\ln(0)$ is undefined [27]. For instance, the Shannon Index for (0, 1, 3) is $0 + 0.34 + 0.21 = 0.55$. The minimum value of the Shannon Index is 0, which indicates that there is no diversity and only one viewpoint was clicked on. When each class is equal the Shannon entropy has the highest value (for three classes, this would be $3 * (-\frac{1}{3}) \ln(\frac{1}{3}) = 1.1$).

Descriptive and Exploratory Measurements. We used these variables to describe our sample and for exploratory analyses, but we did not conduct any conclusive hypothesis tests on them.

- *Demographics* (categorical). We asked participants to state their gender, age group, and level of education from multiple choices. Each of these items includes a "prefer not to say" option.
- *Clicks on Neutral Items* (numerical). In a balanced SERP, the majority of items were neutral. We were specifically interested in whether participants' engagement with search results with a neutral stance is affected by the SERP display condition.
- *Clicking Diversity* (numerical). We logged the clicking behavior of participants during the survey and computed the ratio of pre-search attitude-confirming vs. attitude-opposing search results among the results a user has clicked on. Clicks on neutral search results were not regarded for this variable.
- *Attitude Change* (numerical). In line with previous research [14,15,38], we asked participants to select their attitudes on debated topics before and after the experiments using a seven-point Likert scale ranging from "strongly opposing" to "strongly supporting". The difference between their two answers is then assessed in the analysis.
- *Textual feedback* (free text). We asked participants to provide feedback on the explanations and the task.

Procedure. Participants completed the study in three steps as described below. The survey was conducted on Qualtrics[8], while the interaction with search results occurred on our own server.

Step 1. After agreeing to an informed consent, participants were asked to report their gender, age group, and level of education. Participants were first asked to imagine the following scenario:

> *You and your friend were having a dinner together. Your friend is very passionate about a debated topic and couldn't help sharing his views and ideas with you. After the dinner, you decide to further inform yourself on the topic by conducting a web search.*

Furthermore, participants were asked to state their attitudes concerning each debated topic (see Sect. 3.1; including one attention check for which we specifically instruct participants on what option to select from a Likert scale).

Step 2. We introduced participants to the task and subsequently assigned them to one of the three debated topics (i.e., *atheism*, *intellectual property rights*, and *school uniforms*) depending on their pre-search attitudes and randomly assigned them to one SERP ranking bias condition and one SERP display condition (see Sect. 4.2). Participants were then asked to click on a link leading them to our search platform (i.e., SEPP; see Sect. 3.4). Here, participants could enter as many queries as they want, as long as those queries include their assigned topic term (e.g., `school uniforms pros and cons` for the topic *school uniforms*). Regardless of what or how many queries participants enter, they always received search results from the same pool of 21 available search results relevant to their assigned topic (i.e., seven against, seven neutral, and seven in favor; see Sect. 3.1). With every query that participants entered, they received those search results ranked according to the ranking template associated to their assigned SERP ranking bias condition (see Sect. 3.4) for the first SERP and randomly drawn search results (following the template) for consequent searches.[9] Depending on the SERP display condition participants were assigned to, they could see either plain search results, search results accompanied by stance labels, or search results accompanied by stance labels with additional explanations. Participants were made aware that the search results they were seeing might be biased and that there were limited results. After entering a query, participants were free to explore search results as long as they wish and click on links that lead to the presented web pages, or enter new queries. Users were instructed to return to the Qualtrics survey when they were done searching.

Step 3. Finally, in the questionnaire, we asked participants to report their post-search attitude (towards their assigned topic). Further, we asked them to provide textual feedback on the explanations and the task. We also included another

[8] https://www.qualtrics.com/.

[9] Whenever a user enters a new query, the first SERP (i.e., displaying the top 10 results) will always show search results according to the template, whereas pages 2 and 3 will show the 21 search results relevant to the topic in random order.

attention check to filter out low-quality data in this post-interaction questionnaire. The attention checks consisted of one straightforward question with suggested response options. We excluded the data of participants who failed one or more of the attention checks from data analysis.

Recruitment Methods. In this study, we used Prolific[10] and Qualtrics[11] to manage the participants and design survey workflow, respectively. The workflow of our study, including the informed consent, screening questions, and link to the survey, were all completed on Qualtrics. In the recruitment platform Prolific, we only selected participants with a minimum age of 18 (in compliance with research ethics), and fluent in English, as our dataset only contains English results.

Sample Details. We anticipated to observe medium effects for *SERP display* and *SERP ranking bias* on *clicking diversity* (Cohen's $f = 0.25$). Thus, we determined in an a priori power analysis for a between-subjects ANOVA (see Sect. 4.2) a required sample size of 205 participants, assuming a significance threshold of $\alpha = \frac{0.05}{3} = 0.017$ (testing three hypotheses), a desired power of (1- β) = 0.8 and considering that we tested, depending on the hypothesis, six groups (i.e., three SERP display conditions: *without stance labels, with stance labels, with stance labels and explanation*; and 2 SERP ranking bias conditions: *biased towards the attitude-opposing viewpoint, balanced*) using the software *G*Power* [16]. We aimed for a balanced distribution across topics and conditions.

Participants were required to be older than 18 and with a high proficiency of English (i.e., as reported by *Prolific*). Participants could only participate in our study once. As mentioned above, we excluded participants from data analysis if they did not pass one or more attention checks. We also excluded participants from data analysis if they did not access our search platform at all or if they did not click on any links during their search.

Statistical Analyses. To test our three hypotheses, we conducted Analysis of Variance (ANOVA), looking at the main and interaction effects of the three independent variables (1) the topic, (2) *SERP display* (without stance labels, with stance labels, with stance labels and explanation) and (3) *SERP ranking bias* (biased towards the attitude-opposing viewpoint, balanced) on the *shannon index* (H1a, H1b, H1c). Aiming at a type 1 error probability of $\alpha = 0.05$ and applying Bonferroni correction to correct for multiple testing, we set the significance threshold to $\frac{0.05}{3} = 0.017$. We added *topic* as an additional independent variable to this analysis to control for its potential role as a confounding factor.

In addition to the analyses described above, we conducted posthoc tests (i.e., to analyze pairwise differences) to determine the exact differences and effect size, Bayesian hypothesis tests (i.e., to quantify evidence in favor of null hypotheses), and exploratory analyses (i.e., to note any unforeseen trends in the data) to better understand our results.

[10] https://prolific.co.

[11] https://www.qualtrics.com/.

5 Results

The overarching objective of this research is to investigate the effect of extra visual elements such as stance labels and text explanations on users' interaction behaviours, especially in terms of clicking diversity. In this section, we present the final results of the study with 203 participants and address our research question and hypotheses.

Descriptive Statistics. Prior to analyzing the primary statistics, it is necessary to first examine the demographic data. In general, young people made up most of our participants (see Fig. 3). The educational level data reveal that the majority of participants have completed at least some level of higher education, with a smaller percentage have completed advanced degrees.

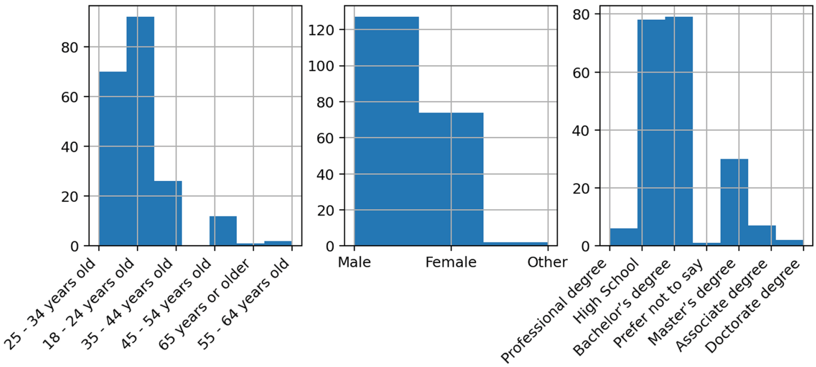


Fig. 3. Demographic information of participants: Bar chart showing the distribution of users' age, gender, and education level.

Participants were roughly equally distributed across the three topics: 70 *atheism*, 66 *intellectual property rights*, and 67 *school uniforms*. Regarding factors such as bias and interface types that were randomized by Qualtrics workflow, their distributions are also balanced, with 102 participants accessed the biased SERPs and 101 accessed balanced SERPs, and 72 users view text-only SERP display, 64 viewed labelled interface and 67 viewed interface with saliency maps. In the pre-survey attitude test, 130 people were granted the access to our survey by expressing rather negative (against) viewpoints towards a specific topic, while only 73 people expressed a positive stance.

Figure 4 shows the diversity of users' clicks as measured by the Shannon index across conditions. For balanced SERPs, the mean Shannon diversity index starts from 0.63 when there are explanation, then it slightly reduces to 0.55 in labelled interface, and drastically drop to 0.24 for the text SERP. For unbalanced SERPs, the trend is similar, from 0.64 to roughly 0.52, but the reduction is less drastic. In

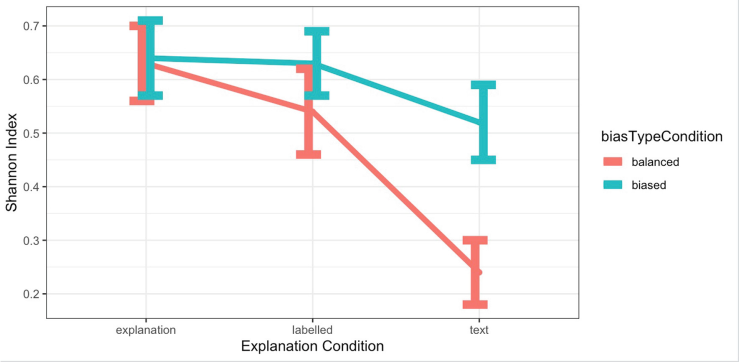


Fig. 4. Shannon Index across SERP Display conditions, split by ranking bias conditions. Error bars represent confidence intervals.

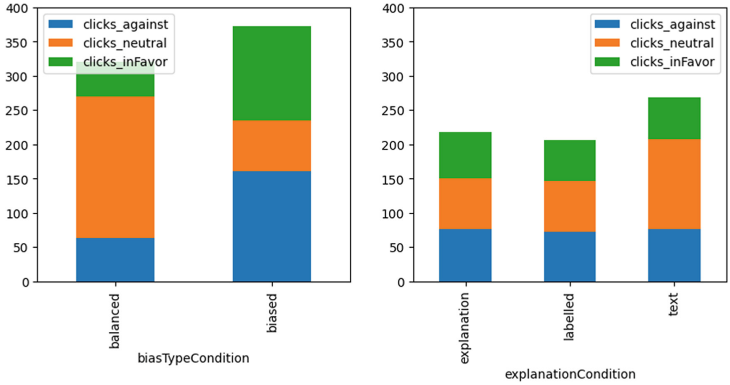


Fig. 5. Number of clicks on items of each stance made by users under different conditions. Left: different bias type, Right: different interface

short, those who interacted with balanced pages overall scored somewhat lower in Shannon diversity.

Figure 5 shows the stacked histogram representing users' click history under different conditions. In the left subplot, we see that *neutral* items attracted more clicks than *favor* and *against* combined in the balanced setting. In the biased setting, however, the number neutral clicks starts to shrink and user start to visit polarized contents more. It is notable that users clicked more *neutral* items in the text-only SERP display. The other two types of interface seem to generate very similar amounts of clicks for each stance.

5.1 Hypothesis Tests

We ran an ANOVA test to examine the relationship between the variable *Shannon index* and other predictor variables, including *explanation condition*, *bias type* and *assigned topic*. Table 3 contains the results of the ANOVA test, including the F-statistic, p-value, and degrees of freedom for each variable and interaction. We also included the assigned topic in the test to control for topic as a potential confounding factor.

Table 3. ANOVA results for the Shannon index and independent variables including the explanation level, bias type, the interaction between explanation and bias type, and the topic the user holds strong opinion towards. Under our reduced significance threshold, only the *Explanation* effect is significant.

Variable	Df	Sum Sq	Mean Sq	F	*p*
SERP Ranking Bias	1	0.81	0.8146	4.911	.027
SERP Display	2	2.54	1.2712	7.664	<.001
Topic	2	0.03	0.0158	0.095	.909
SERP RB:SERP Display	2	0.65	0.3256	1.963	.143
Residuals	195	32.34	0.1659		

H1a: Users who are exposed to viewpoint-biased search results interact with less diverse results than users who are exposed to balanced search results. To test this hypothesis, we examine the Shannon diversity of clicks made by users who viewed viewpoint-biased search engine result pages (SERPs) to those who did not. Figure 4 suggests that users who were exposed to more balanced search results clicked on somewhat less diverse content. However, in the ANOVA summary (Table 3), we see that the influence of bias type ($F = 4.911, df = 1, p = .027$, Cohen's $f = 0.16$) is not statistically significant, given the significance threshold of .017. We thus do not find any conclusive evidence for a difference in the diversity of clicked results between bias conditions.

H1b: Users who are exposed to search results with (1) stance labels or (2) stance labels with explanations for each search result interact with more diverse content than users who are exposed to regular search results. Regarding H1b, compare users' clicks in different SERP displays. The results of the ANOVA shows a significant effect of explanation condition on clicking diversity ($F = 7.664, df = 2, p < .001$, Cohen's $f = 0.28$). This suggests that there is a difference in the diversity of content interacted between users who were exposed to (1) plain search results, (2) search results with stance labels, or (3) search results with stance labels and also explanations. We conducted a pairwise Tukey test to determine whether there are significant differences between SERP display levels (text, label, explanation). We found significant differences between `text-label` and `text-explanation` (with adjusted p-values of .0008 and .012, respectively), suggesting that both stance labels and stance labels with explanations led to more diverse clicks compared to the regular SERPs. The p-value of `labelled-explanation` group is .723, indicating that there may be no difference in Shannon diversity between the *labelled* and *explanation* group.

H1c: Users who are exposed to search results with (1) stance labels or (2) stance labels with explanations are less susceptible to the effect of viewpoint biases in search results on clicking diversity. This hypothesis concerns the interaction effect between SERP display and bias types on the Shannon index. Our ANOVA (Table 3) did not reveal any evidence for such an interaction effect ($F = 1.963, df = 2, p = .143$, Cohen's $f = 0.02$). In other words, when users are looking at different types of SERP displays but are exposed to the same level of viewpoint bias, our results do not contain evidence that users will click more diverse items because explanations and labels are visible.

5.2 Exploratory Analysis

To further understand our results, we also conducted exploratory analysis. Note that these analyses were not preregistered.

Clicks on Neutral Items. To better understand why the search results were more diverse in the labeled and explanation conditions, we looked closer at the distribution across the three viewpoints. In Fig. 5, we can already see a larger number of neutral results for the text-only condition. Furthermore, in every SERP display, users who were exposed to a balanced page clicked on more neutral items on average. This may be due to their tendency to click on highly-ranked neutral items while being unaware of the stance.

We conducted an exploratory ANOVA to investigate the effects of biases and interfaces on the number of clicks on neutral items. Table 4 lists the test results.

Table 4. Results of the ANOVA analysis on for number of neutral clicks.

Variable	Df	Sum Sq	Mean Sq	F	p
SERP Ranking Bias	1	84.3	84.30	44.308	<.001
SERP Display	2	23.1	11.54	6.067	.003
SERP RB:SERP Display	2	15.6	7.79	4.096	.018
Residuals	197	374.8	1.90		

We can observe that both explanation condition and bias condition had main effects on click diversity. Also their interaction is significant. This means that users may click on more neutral results in the balanced condition, and especially so when SERPs do not contain any stance labels or explanations.

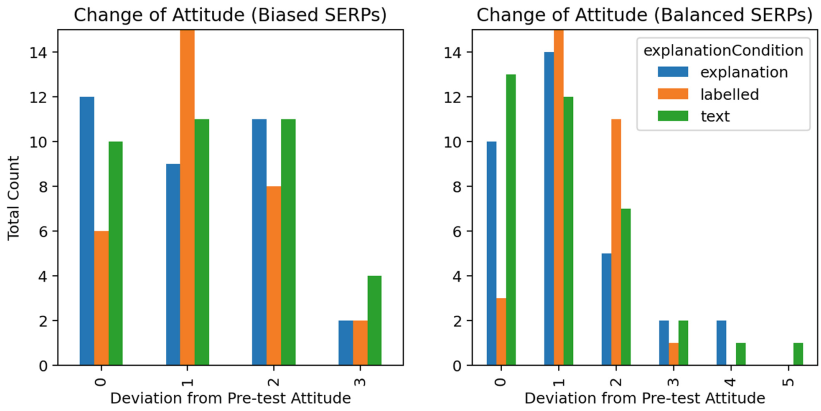


Fig. 6. Histogram visualizing the difference between pre- and post-survey attitudes. Each group of bins indicates the number of people in a bias type setting who changed their viewpoints towards their assigned topic by a certain number of points. The X-axis indicates the normalized difference of attitude between the pre- and post-survey answers measured by a 7-point Likert Scale.

Table 5. Absolute values of attitude change per condition.

Bias Type	SERP display	Attitude Change (Mean)	Attitude Change (Median)	Skewness
Balanced	Text-only	1.14	1.00	1.31
	Labelled	1.30	1.00	0.15
	Explanation	1.15	1.00	1.10
Biased	Text-only	1.25	1.00	0.20
	Labelled	1.19	1.00	0.35
	Explanation	1.09	1.00	0.25

Attitude Change. Previous research showed that mildly opinionated users' attitude change can differ across levels of ranking bias [3,8,14,15,33]. However, this effect has, to the best of our knowledge, not yet been shown for strongly opinionated users such as in our study. We intended to measure the attitude change by subtracting the post-search viewpoint t_1 from the pre-search viewpoint t_0, and thus the difference would range from -6 to 6. We summarize users' attitude change in Fig. 6 and Table 5.

From Fig. 6, we observe that only a few participants developed large attitude changes (absolute value > 3) after the survey. In Table 5, the mean absolute attitude change ranges from 1.0 to 1.3. The median statistics appear to be very stable; under all conditions, 1 is always the central value. We performed an exploratory ANOVA on the absolute *attitude change* variable (Table 6) but do not find any convincing evidence for an effect of any independent variables on attitude change in our scenario, i.e., apart from a potential role of the topic ($F = 4.132, df = 2, p = .017$) that would need to be further investigated.

Table 6. ANOVA Results: absolute value of users' attitude change

Variable	Df	Sum Sq	Mean Sq	F	p
SERP Ranking Bias	1	0.01	0.013	0.014	.907
SERP Display	2	0.56	0.282	0.296	.744
Topic	2	7.88	3.938	4.132	.017
SERP RB:SERP Display	2	0.55	0.277	0.290	.748
Residuals	195	185.88	0.953		

5.3 Qualitative Feedback

After excluding feedback with fewer than four characters like "no", "nope" etc., there are 59 substantial comments. Positive feedback indicated that participants perceived the search results to be diverse: *"I thought the search engine was quite varied as it provided different types of sources"*. Other participants perceived the viewpoint labels to be accurate: *"The search engine was very useful, and the classification of the information was accurate."* At least some participants were able to perceive the bias in the search result pages: *"the results against atheism were promoted to the top results of the search, it was not an impartial search result"*. While we aimed to mimic a real search engine, some participants may have realized that the results were reshuffled: *"It appears the results of the search were similar or the same in each new search. I'm not sure the words used to determine the stance of the page were appropriate."*

6 Discussion and Limitations

We found a significant effect of *SERP display* on clicking diversity. That is, participants who were exposed to viewpoint labels or viewpoints labels with explanations consumed more diverse results than plain text search. While both non-text SERP displays affect users' click diversity over text-only SERP displays, we cannot observe any additional effect from the SERP display with labels and explanations over the label-only SERP display. Our results suggest that this difference can be explained by a predominance of clicks on the neutral stance in the text-only condition. This, alongside the qualitative comments, suggests that participants trusted and used the labels and explanations to inform themselves diversely. Contrary to our expectations, we did not find evidence for a difference between *biased and balanced search results*, in terms of click diversity. Furthermore, exploratory analyses revealed that users exposed to balanced SERPs clicked on more neutral items.

We further found no exploratory evidence that intervention types (bias of search results; explanation level) affect participants' viewpoints. This later result is not necessarily surprising, given that the participants in this study held strong opinions on the topic. We also would not expect a change because the task given to participants was formulated as low-stakes and open-ended (inform yourself about the topic after speaking to an opinionated friend).

Limitations and Future work. Our study has at least two important limitations. First, each search result in the data set we considered only had one overall viewpoint label, and this was limited to *against*, *neutral*, and *in favor*. This allowed for scalability but does not reflect the full nuances of online search results. For example, an essay or blog post could express highly diverse perspectives but still receive a positive stance label if that is its overall conclusion. Secondly, our study carefully aimed to balance a controlled environment with maintaining ecological validity. Despite this, it is possible that participants recognized that this was not a true search engine when issuing new queries. Similarly, the selection of templates allowed for some structured reshuffling of results. However, this also meant that strong opinions (contrary to the active user) were always at the top of biased search results. For the balanced condition there were also more neutral results. This likely contributed to the higher diversity of search results selected in the biased condition. In addition, ranking fairness metrics could have been used for the ranked lists, which could have led to slightly different results. Further work is required to disentangle the relationship between position bias and confirmation bias, and to replicate the study with different templates.

7 Conclusion

In this paper, we studied the impact of stance labels and explanations of these labels for search on disputed topics. We found that stance labels and explanations led to a more diverse search result consumption. However, we cannot conclude that explanations have an extra effect in addition to labels on user's click diversity. Whether consuming diverse or more neutral results is preferable is in itself a debated topic. Backfire effects, where users become even more invested in their pre-existing beliefs, are possible when users consume strongly opposing views [32]. Greater diversity can further induce wrong perceptions of present evidence: for example, portraying climate change deniers and believers equally can give the impression that climate change is an open issue and thus may be worse than indeed weighing the evidence on both sides [11]. How much stance diversity is ideal can thus depend on individual user traits [30] and may lie somewhere between the extremes [7]. While these are domains and a context where confirmation biases are expected to be large, similar cognitive biases are likely to occur in other decision-making tasks where XAI is used. Further work is needed to catalogue different scenarios of opinion formation in different domains where confirmation bias may be present.

Acknowledgments. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 860621.

References

1. Aldayel, A., Magdy, W.: Your stance is exposed! Analysing possible factors for stance detection on social media. In: Proceedings of the ACM on Human-Computer Interaction, vol. 3(CSCW), pp. 1–20 (2019)
2. Aldayel, A., Magdy, W.: Stance detection on social media: state of the art and trends. Inf. Process. Manage. **58**(4), 102597 (2021). https://doi.org/10.1016/j.ipm.2021.102597
3. 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 (2014). https://doi.org/10.2196/jmir.2642. http://www.jmir.org/2014/4/e100/
4. Allaway, E., McKeown, K.: Zero-shot stance detection: a dataset and model using generalized topic representations. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 8913–8931. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.emnlp-main.717. https://aclanthology.org/2020.emnlp-main.717
5. Augenstein, I., Rocktäschel, T., Vlachos, A., Bontcheva, K.: Stance detection with bidirectional conditional encoding. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 876–885. Association for Computing Machinery, New York, NY, USA (2016)
6. Azzopardi, L.: Cognitive biases in search: a review and reflection of cognitive biases in information retrieval (2021)
7. Bail, C.A., et al.: Exposure to opposing views on social media can increase political polarization. Proc. Natl. Acad. Sci. **115**(37), 9216–9221 (2018)
8. Bink, M., Schwarz, S., Draws, T., Elsweiler, D.: Investigating the influence of featured snippets on user attitudes. In: ACM SIGIR Conference on Human Information Interaction and Retrieval. CHIIR 2023, ACM, New York, NY, USA (2023). https://doi.org/10.1145/3576840.3578323
9. 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. ACM, Regensburg Germany (2022). https://doi.org/10.1145/3498366.3505766
10. Chen, S., Xiao, L., Kumar, A.: Spread of misinformation on social media: what contributes to it and how to combat it. Comput. Hum. Behav. **141**, 107643 (2023). https://doi.org/10.1016/j.chb.2022.107643
11. Cushion, S., Thomas, R.: From quantitative precision to qualitative judgements: professional perspectives about the impartiality of television news during the 2015 UK general election. Journalism **20**(3), 392–409 (2019)
12. Draws, T., et al.: Explainable cross-topic stance detection for search results. In: CHIIR (2023)
13. Draws, T., et al.: Viewpoint diversity in search results. In: Kamps, J., et al. (eds.) ECIR 2023. LNCS, vol. 13980, pp. 279–297. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-28244-7_18
14. 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. SIGIR 2021, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3404835.3462851. https://dl.acm.org/doi/10.1145/3404835.3462851

15. Epstein, R., Robertson, R.E.: The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. Proc. Natl. Acad. Sci. **112**(33), E4512–E4521 (2015). https://doi.org/10.1073/pnas.1419828112. http://www.pnas.org/lookup/doi/10.1073/pnas.1419828112
16. Faul, F., Erdfelder, E., Buchner, A., Lang, A.G.: Statistical power analyses using g* power 3.1: tests for correlation and regression analyses. Behav. Res. Meth. **41**(4), 1149–1160 (2009)
17. Feldhus, N., Hennig, L., Nasert, M.D., Ebert, C., Schwarzenberg, R., Möller, S.: Constructing natural language explanations via saliency map verbalization. arXiv preprint arXiv:2210.07222 (2022)
18. Gezici, G., Lipani, A., Saygin, Y., Yilmaz, E.: Evaluation metrics for measuring bias in search engine results. Inf. Retrieval J. **24**(2), 85–113 (2021). https://doi.org/10.1007/s10791-020-09386-w
19. Gohel, P., Singh, P., Mohanty, M.: Explainable AI: current status and future directions. arXiv preprint arXiv:2107.07045 (2021)
20. Hanselowski, A., et al.: A retrospective analysis of the fake news challenge stance-detection task. In: Proceedings of the 27th International Conference on Computational Linguistics, pp. 1859–1874. Association for Computational Linguistics, Santa Fe, New Mexico, USA (2018). https://aclanthology.org/C18-1158
21. Hardalov, M., Arora, A., Nakov, P., Augenstein, I.: Few-shot cross-lingual stance detection with sentiment-based pre-training. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 10729–10737. AAAI (2022)
22. Jin, W., Carpendale, S., Hamarneh, G., Gromala, D.: Bridging AI developers and end users: an end-user-centred explainable AI taxonomy and visual vocabularies. In: Proceedings of the IEEE Visualization, Vancouver, BC, Canada, pp. 20–25 (2019)
23. Kaiser, B., Wei, J., Lucherini, E., Lee, K., Matias, J.N., Mayer, J.: Adapting security warnings to counter online disinformation. In: 30th USENIX Security Symposium (USENIX Security 2021), pp. 1163–1180 (2021)
24. Küçük, D., Can, F.: Stance detection: a survey. ACM Comput. Surv. **53**(1), 1–37 (2021). https://doi.org/10.1145/3369026
25. Leonhardt, J., Rudra, K., Anand, A.: Extractive explanations for interpretable text ranking. ACM Trans. Inf. Syst. **41**(4), 1–31 (2023). https://doi.org/10.1145/3576924
26. Lyu, L., Anand, A.: Listwise explanations for ranking models using multiple explainers. In: Kamps, J., et al. (eds.) ECIR 2023. LNCS, vol. 13980, pp. 653–668. Springer, Heidelberg (2023). https://doi.org/10.1007/978-3-031-28244-7_41
27. MacKay, D.J.: Information Theory, Inference and Learning Algorithms. Cambridge University Press, Cambridge (2003)
28. Madsen, A., Reddy, S., Chandar, S.: Post-hoc interpretability for neural NLP: a survey. ACM Comput. Surv. **55**(8), 1–42 (2022)
29. Mena, P.: Cleaning up social media: the effect of warning labels on likelihood of sharing false news on Facebook. Policy Internet **12**, 165–183 (2020). https://doi.org/10.1002/poi3.214
30. 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 (2010)
31. Nickerson, R.S.: Confirmation bias: a ubiquitous phenomenon in many guises. Rev. Gen. Psychol. **2**, 175–220 (1998)
32. Nyhan, B., Reifler, J.: When corrections fail: the persistence of political misperceptions. Polit. Behav. **32**(2), 303–330 (2010)

33. 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 (2017). https://doi.org/10.1145/3121050.3121074
34. Puschmann, C.: Beyond the bubble: assessing the diversity of political search results. Digit. Journal. **7**(6), 824–843 (2019). https://doi.org/10.1080/21670811.2018.1539626
35. Putra, S.R., Moraes, F., Hauff, C.: SearchX: empowering collaborative search research. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, pp. 1265–1268 (2018)
36. Reuver, M., Verberne, S., Morante, R., Fokkens, A.: Is stance detection topic-independent and cross-topic generalizable? - a reproduction study. In: Proceedings of the 8th Workshop on Argument Mining, pp. 46–56. Association for Computational Linguistics, Punta Cana, Dominican Republic (2021). https://doi.org/10.18653/v1/2021.argmining-1.5
37. Ribeiro, M.T., Singh, S., Guestrin, C.:"Why should i trust you?": explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016)
38. Rieger, A., Draws, T., Tintarev, N., Theune, M.: 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. HT 2021, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3465336.3475101
39. Roy, A., Fafalios, P., Ekbal, A., Zhu, X., Dietze, S.: Exploiting stance hierarchies for cost-sensitive stance detection of web documents. J. Intell. Inf. Syst. **58**(1), 1–19 (2022). https://doi.org/10.1007/s10844-021-00642-z
40. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. ArXiv abs/1910.01108 (2019)
41. Sepúlveda-Torres, R., Vicente, M., Saquete, E., Lloret, E., Palomar, M.: Exploring summarization to enhance headline stance detection. In: Métais, E., Meziane, F., Horacek, H., Kapetanios, E. (eds.) NLDB 2021. LNCS, vol. 12801, pp. 243–254. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-80599-9_22
42. Silalahi, S., Ahmad, T., Studiawan, H.: Named entity recognition for drone forensic using BERT and distilBERT. In: 2022 International Conference on Data Science and Its Applications (ICoDSA), pp. 53–58. IEEE (2022)
43. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**(1), 1929–1958 (2014)
44. Staliūnaitė, I., Iacobacci, I.: Compositional and lexical semantics in RoBERTa, BERT and distilBERT: a case study on COQA. arXiv preprint arXiv:2009.08257 (2020)
45. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: International Conference on Machine Learning (2017)
46. Tong, J., Wang, Z., Rui, X: A multimodel-based deep learning framework for short text multiclass classification with the imbalanced and extremely small data set. Comput. Intell. Neurosci. **2022** (2022)
47. 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 (2013). https://doi.org/10.1145/2484028.2484053

48. Wolf, T., et al.: Transformers: state-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 38–45. Association for Computational Linguistics (2020). https://www.aclweb.org/anthology/2020.emnlp-demos.6
49. Xu, C., Paris, C., Nepal, S., Sparks, R.: Cross-target stance classification with self-attention networks. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (vol. 2: Short Papers), pp. 778–783. Association for Computational Linguistics, Melbourne, Australia (2018). https://doi.org/10.18653/v1/P18-2123. https://aclanthology.org/P18-2123
50. 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 (2017). https://doi.org/10.1145/3085504.3085526
51. Ying, X.: An overview of overfitting and its solutions. In: Journal of Physics: Conference Series, vol. 1168, p. 022022. IOP Publishing (2019)
52. Yu, P., Rahimi, R., Allan, J.: Towards explainable search results: a listwise explanation generator. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 669–680. SIGIR 2022, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3477495.3532067
53. Zehlike, M., Yang, K., Stoyanovich, J.: Fairness in ranking: a survey. arXiv:2103.14000 [cs] (2021). http://arxiv.org/abs/2103.14000