Quantifying Political Bias in News Articles

GIZEM GEZICI, Huawei Turkey R&D Center, Turkey

Search bias analysis is getting more attention in recent years since search results could affect In this work, we aim to establish an automated model for evaluating ideological bias in online news articles. The dataset is composed of news articles in search results as well as the newspaper articles. The current automated model results show that model capability is not sufficient to be exploited for annotating the documents automatically, thereby computing bias in search results.

Additional Key Words and Phrases: Bias evaluation, Fair ranking, Political bias, Web search

ACM Reference Format:
Gizem Gezici. 2022. Quantifying Political Bias in News Articles. 1, 1 (October 2022), 9 pages. https://doi.org/10.1145/nnnnnn.nnnnnnn

1 INTRODUCTION

Search engines are ubiquitous. As reported by SmartSights [3], in 2017 46.8% of the world population accessed the internet and by 2021, the number is expected to reach 53.7%. Currently, on average 3.5 billion Google searches are done per day [2]. These statistics advocate that search systems are “gatekeepers to the Web” for many people [6]. As information seekers search the Web more, they are also more influenced by Search Engine Result Pages (SERPs) and their influence -negative included- potentially become visible in a wide range of areas. However, as with all software-based systems, search platforms do not lack of human influence, thereby they may suffer from embedded bias, i.e., corpus or algorithmic bias. Experimental studies suggest that particular information types or sources might be retrieved more or less, or might not be well represented [4]. For instance, during the elections, it is known that people issue repeated queries on the Web about political candidates and events such as “democratic debate”, “Donald Trump” and “climate change” [10]. Epstein and Robertson [7] claim that SERPs returned in response to these queries may influence the voting decisions of the users and report that manipulated search rankings can change the voting preferences of undecided individuals at least by 20%. As individuals rely to a greater extent on the SERPs for their decision making, there is a thriving demand for these systems explainable.

Empirical evidence has shown that individuals trust more sources ranked at higher positions in SERPs, but the ranking criteria may rather depend more on user satisfaction than the factual information, which jeopardizes the phenomenon of providing reliable information in exchange for satisfying users [4]. In spite of the given research findings, the majority of online users tend to believe that search engines provide neutral results, i.e., serving only as facilitators in accessing information on the Web due to their automated operations [9]. However, this romanticised view of search platforms does not reflect reality and there seems to be a growing skepticism related to objectivity and credibility of these platforms. To illustrate that, a recent dispute between the U.S. President Donald Trump and Google can be given, where Mr. Trump accused Google of
presenting only negative news about him when his name is searched. Google refuted this claim by saying that: “When users type queries into the Google Search bar, our goal is to make sure they receive the most relevant answers in a matter of seconds” and “Search is not used to set a political agenda and we don’t bias our results toward any political ideology”\(^1\). In this work, we hope to shed some light on that debate, by not specifically examining queries regarding Donald Trump but by fulfilling an in depth analysis of retrieved search answers to a broad set of queries related to controversial topics based on concrete evaluation measures.

Bias implies undue emphasis. For a retrieval system, it can be defined as the balance and representativeness of Web documents retrieved from a database for a set of queries \([11]\). When a user issues a query to a search engine, documents from different sources are collected, ranked, and presented to the user. Assume that a user searches for 2016 presidential election and the top-\(n\) ranked results are displayed. In such a search scenario, the retrieved results may exaggerate or downplay particular perspectives and thereby provide an unbalanced picture of the given query as claimed by Mr. Trump, though without any scientific support. Hence, the potential undue inclusion or exclusion of specific perspectives in the retrieved results lead to bias \([11]\). Note that the existence of bias is different from relevance. In the presented scenario, even though the retrieved list of documents may all be judged as relevant with respect to the given query, if the selection of the documents is skewed or slanted, i.e., emphasizing one perspective over another, then the corresponding search engine is biased due to an imbalanced representation of the perspectives towards the query’s topic. Bias is especially important if the query topic is controversial having opposing perspectives as described in the given scenario. The bias in SERPs can be used by search engines to inform their users about the bias by making themselves more accountable which is one of the crucial attributes that a retrieval system should possess \([4]\).

In this work, we focus on the SERPs coming from the news sources and investigate two major search engines (Bing and Google) in terms of political bias. Our analysis has mainly three sides where we evaluate the political bias of the search engines separately, then make a comparison among them as well as track the source of this bias in the SERPs. The bias may come from the data, which may contain biases (input bias) or the search algorithm, which contains sophisticated features (algorithmic bias). In the scope of this work, we concentrate on the input bias that is intrinsically embedded in the data itself.

In short, we aim to answer the following research questions:

**RQ1:** On a conservative-to-liberal scale, do search engines return politically biased SERPs in response to queries related to controversial topics?

**RQ2:** Are search engines significantly different from each other towards controversial topics?

**RQ3:** Does the source of bias come from the input data?

We address these research questions for controversial topics representing a broad range of issues in SERPs of Google and Bing through content analysis, i.e. analysing the textual content of the retrieved documents. We focus on content bias and describe our SERP politically content bias quantification framework in which we propose three different measures of bias based on common Information Retrieval (IR) utility-based evaluation measures: Precision at cut-off (P@n), Rank Biased Precision (RBP), and Discounted Cumulative Gain at cut-off (DCG@n). While the first measure quantifies bias considering only a weak ranking criterion, i.e. the first \(n\) returned documents as in SERPs, the other two measures incorporate stronger ranking bias.

In order to answer RQ1, we measure the degree of deviation of the ranked SERPs from an ideal distribution where different political perspectives are equally likely to appear \([11]\). To detect

---

\(^1\)https://www.reuters.com/article/us-usa-trump-tech-alphabet/google-responds-to-trump-says-no-political-motive-in-search-results-idUSKCN1LD1QP

, Vol. 1, No. 1, Article . Publication date: October 2022.
political bias which results from the imbalanced representation of politically different points of view, we label the documents’ political perspectives with classification and use these labels for bias evaluation. To address RQ2, we compare the political bias in the SERPs of the two search engines to see if they show similar bias for the corresponding controversial topics. To answer RQ3, we measure the political bias in a similar manner as we do to answer the RQ1. The only difference is that this time we measure the bias in the whole corpus, i.e., all retrieved SERPs from Google and Bing, instead of the partial of it, as top $n$ documents. Then, we compare the bias of the whole corpus with the partial bias that we compute for the first part of our analysis to check if they are consistent, or not.

Our main contributions in this work can be summarized as follows:

1. We propose a novel generalizable search bias evaluation framework to measure the bias in the SERPs.
2. Our fairness-aware set of measures are explainable utility-based IR measures which take into account of relevance as well, while evaluating bias in ranked results.
3. We apply our framework to compare the relative bias in the SERPs content of two search engines for controversial queries. To the best of our knowledge, this is one of the first works to compare the two search engines in terms of content bias.
4. We utilize the framework to track the source of bias in SERPs, i.e., checking if the existing bias comes from the corpus or not.

Lastly, a preliminary version of the present framework has been published earlier [8]. This work expands and surpasses the previous endeavor mainly in the following points: (i) We used a deep learning framework that is widely-used in NLP tasks for classification, i.e., to predict the political perspectives of the retrieved documents from the two popular search engines, instead of obtaining these labels via crowdsourcing. (ii) We also crawled a different dataset of articles from the newspapers’ websites to fine-tune the deep learning model specifically for the news articles, thereby achieve a better classification model in perspective detection. (iii) Finally, we included a new investigation on the source of bias in our experiments through examining if the measured bias in the top-n documents comes from the data.

The remainder of the paper is structured as follows. In Section 2 we present the related work. The search bias evaluation framework is developed in Section 3. In Section 4 we detail the experimental setup, and present the results.

2 BACKGROUND & RELATED WORK

Studies indicate that online IR platforms have influenced public by causing an increase in polarization; for instance it has been shown that social media accounts with non-Western names are more likely to be indicated as fraudulent since most probably the system has been trained on Western names. Therefore, proposing new metrics can be useful in evaluating and understanding such ethical issues in IR systems [4]. Regarding this assertion, there have been a growing interest in quantifying bias in the recent years.

3 POLITICAL BIAS EVALUATION FRAMEWORK

In this section, to answer the research questions asserted in the introduction, we propose our political bias evaluation framework. The first question focuses on detecting political bias (if exists) in SERPs and the second one examines if the search engines show the same level of bias. On the other hand, the third question investigates if the source of bias comes from the data. In this framework, we initially identify political perspectives of the news articles automatically via classification. Then,
we present the measures of bias and the proposed protocol to identify political bias and investigate the source of bias in SERPs.

3.1 Preliminaries

Our first goal is to detect political bias with respect to the distribution of political perspectives expressed in the contents of the SERPs. Let $S$ be the set of search engines and $Q$ be the set of queries about controversial topics. When a query $q \in Q$ is issued to a search engine $s \in S$, the search engine $s$ returns a SERP $r$. We define the political perspective of the $i$-th retrieved document $r_i$ with respect to $q$ as $j(r_i)$. A political perspective can have the following values: conservative, liberal, both or neither.

A document political perspective can be:

- **conservative (●)** when the document content is in favour of the conservative political agenda.
- **liberal (●)** when the document content is in favour of the liberal political agenda.
- **both or neither (○)** when the document content is in favour of both or neither perspectives.

For our analyses, we deliberately use controversial topics such as abortion, medical marijuana, gay marriage, and Cuba embargo which contain opposing viewpoints since complicated concepts concerning the identity, religion, or political leaning are the actual points where search engines are more likely to provide biased results [12]. Our second goal is related to the first one as we compare the search engines in terms of the political bias that they show in their SERPs.

Our third goal is to find the source of political bias by checking if the bias comes from the data (input bias). We do this by measuring the bias also in the whole corpus of the SERPs in addition to the partial dataset, i.e. top 10 or 100 documents of the corpus, and comparing the corpus bias (if exists) with the partial one to see if they are consistent or not.

3.2 Measures of Bias

On the basis of the previously definition of bias given in Section 1, it can be quantified by measuring the degree of deviation of the distribution of documents from the ideal one. To give a generic definition of an ideal list in the field of bias evaluation poses problems; but in the scope of this work, we can assess the existence of political bias in a ranked list retrieved by a search engine if the presented information significantly deviates from true likelihoods [14]. Using this definition reversely, we can adopt the assumption that the ideal list is the one that minimizes the difference between two opposing political perspectives, which we indicate here as ● and ○.

3.2.1 Bias Score of a SERP. Formally, we measure the political bias in a SERP $r$ as follows:

$$
\beta_f(r) = f_\bullet(r) - f_\circ(r),
$$

where $f$ is a function that measures the likelihood of $r$ in satisfying the information need of the user about the view ● and the view ○.

3.2.2 Mean Bias (MB). Before defining $f$, from Eq. (3.2.1), we define the mean bias of a search engine $s$ as:

$$
MB_f(s, Q) = \frac{1}{|Q|} \sum_{q \in Q} \beta_f(s(q)).
$$

An unbiased search engine would produce a mean bias of 0. A limitation of MB is that if a search engine is biased towards the ● perspective on one topic and bias towards the ○ perspective on another topic, these two contributions will cancel each other out. In order to avoid this limitation we also define another metric.
3.2.3 Mean Absolute Bias (MAB). The mean absolute bias, which consists in taking the absolute value of the bias for each \( r \). Formally, this is defined as following:

\[
\text{MAB}_f(s, Q) = \frac{1}{|Q|} \sum_{q \in Q} |\beta_f(s(q))|.
\]

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

3.2.4 Retrieval Evaluation Measures. In IR the likelihood of \( r \) in satisfying the information need of users is measured via retrieval evaluation measures. Among these measures we selected 3 utility-based evaluation measures. This class of evaluation measures quantify \( r \) in terms of its worth to the user and are normally computed as a sum of the information gain summed over the relevant documents retrieved by \( r \).

The 3 IR evaluation measures used in the following experiments are: P@n, RBP, and DCG@n.

**P@n:** P@n for the \( \Theta \) view is formalized in [8]. However, differently from the previous definition of \( j \) where the only possible outcomes are \( g_1 \) and \( g_2 \), here \( j \) can return any of the label associated to a political perspective (\( \bigcirc \), \( \bigotimes \), and \( \bigcirc \)). Hence, only conservative and liberal documents, that is relevant to the topic, are taken into account, since \( j(r_i) \) returns both or neither when otherwise.

Replacing P@n in Eq. (3.2.1) we obtain the first measure of bias:

\[
\beta_{P@n}(r) = \frac{1}{n} \sum_{i=1}^{n} ([j(r_i) = \bigotimes] - [j(r_i) = \bigotimes]).
\]

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

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

\[
\text{RBP}_\Theta = (1 - p) \sum_{i=1}^{n} p^{i-1} [j(r_i) = \bigotimes].
\]

Substituting Eq. (3.2.4) to Eq. (3.2.1) we obtain:

\[
\beta_{\text{RBP}}(r) = (1 - p) \sum_{i=1}^{n} p^{i-1} ([j(r_i) = \bigotimes] - [j(r_i) = \bigotimes]).
\]

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

\[
\text{DCG}_\Theta@n = \frac{1}{\log(n+1)} \sum_{i=1}^{n} [j(r_i) = \bigotimes].
\]

Substituting Eq. (3.2.4) to Eq. (3.2.1) we obtain:

\[
\beta_{\text{DCG}@n}(r) = \frac{1}{\log(n+1)} \left( [j(r_i) = \bigotimes] - [j(r_i) = \bigotimes] \right)
\]

Since we are evaluating web-users, for P@n and DCG@n we set \( n = 10 \) and for RBP we set \( p = 0.8 \).

This last formulation (Eq. ), although it looks similar to the rND measure, it does not suffer from the first four limitations introduced in Section 2. In particular all these presented measures of bias:
3.3 Quantifying Political Bias

Using the previously defined measures of bias in Section 3.2 we quantify the political bias of the two search engines, Bing and Google. Then we compare them in terms of the measured bias. Now, we describe each step of the proposed mechanism used to quantify political bias in SERPs.

**Dataset Construction.** After having crawled all the SERPs for both search engines and all queries $Q$, for each returned document we obtain the political perspective of the document with respect to the corresponding query in the top 10 documents and the whole corpus. Both is done automatically via classification since we already have a dataset labelled by crowdsourcing for training.

**Bias Evaluation.** We compute the bias measures of every SERP with all three IR-based measures of bias: $P@n$, RBP, and DCG@$n$ for the first 10 documents and only $P@n$ for the whole corpus, because the other measures of bias do not provide meaningful results due to the rank information used in these measures. We then aggregate the results using the two measures of bias, MB and MAB.

**Statistical Analysis.** To identify whether the bias measured is not rooted from randomness, for the bias measure MAB, we compute a one-sample t-test: the null hypothesis is that no difference exists and that the true mean is equal to zero. If this hypothesis is rejected, this means that there is a significant difference; we claim that the evaluated search engine is biased. Then, we make a comparison in the bias difference measured across the two search
engines using a two-tailed paired t-test: the null hypothesis is that the difference between
the two true means is equal to zero. If this hypothesis is rejected, this means that there is a
significant difference so we claim that there is a difference in bias between the two search
engines.

4 EXPERIMENTAL SETUP

4.1 Dataset

In this work, we used two datasets for political bias evaluation. The first dataset contains SERPs
crawled from two search engines, Google and Bing. The second dataset, on the other hand is
composed of the articles published in the known newspapers.

News Articles in SERPs. We obtained the controversial queries issued for searching from Pro-
Cong.org [13] and applied some filtering steps on the initial query set. After filtering, the
final query set size became 57. We submitted each query in the final query set to the US
News search engines of Google and Bing using a US proxy. Then, we extracted the whole
corpus returned by both engines in response to all the queries in the set. Note that the data
collection process was done in a controlled environment such that the queries are sent to
the search engines at the same time. For more details about the selection of the queries
and crawling the SERPs, please refer to our previous work [8]. After having crawled all the
SERPs returned from both engines and extracted their contents, we annotated the top 10
documents. We obtained the stance label of each document with respect to the queries via
crowdsourcing. To label the political perspective of queries, we also used crowdsourcing. To
obtain the political perspective of documents, we transformed the stance labels into political
perspectives based on the perspective of their corresponding queries. The details about our crowdsourcing campaigns as well as the transformation process can also be found in [8].

**News Articles in Newspapers.** To enrich the pre-trained BERT model on perspective detection task even more, we crawled the article contents of the selected newspapers. To choose the newspapers to be crawled, we prepared a list by utilizing the media bias chart of AllSides.com All [1]. AllSides.com is a US-based online website that shares information and ideas from all sides of the political spectrum with the purpose of fighting filter bubbles and polarization. The organization has researched since 2012, uses multiple methodologies to rate media bias and even achieved a patent on this area. Furthermore, AllSides only focuses on American sources and its media bias rating scale is based on American politics for the online news articles. For all these reasons, we chose the media bias chart of AllSides to determine the political perspective of newspapers which is displayed in Figure 2. Subsequently, we crawled the textual contents of the news articles of the corresponding newspapers in the chart by using a Python library called feedparser that parses feeds in all known formats such as Atom, RSS, and RDF. Then, we labelled the crawled news articles by weak supervision such that labelling the articles of left/lean-left newspaper as liberal, right/lean-right as conservative, and center as neutral. We further used this newspaper dataset to fine-tune the BERT model after the first fine-tuning done with the training data by learning the topic keywords and semantic relations in the articles, thereby detecting the political perspectives more accurately.

4.2 **Classification**

To answer the first research question, we used the political perspective labels of the top 10 documents in the first dataset previously obtained via crowdsourcing. In the scope of this work, we used the labelled dataset to obtain those perspectives automatically. For automatic perspective detection, we used pre-trained Bidirectional Encoder Representations from Transformers (BERT) model which is proposed by Google that presents state-of-the-art results in a wide variety of NLP tasks [5]. BERT’s key difference from other deep models is having an attention mechanism called Transformer that applies bidirectional training as opposed to directional models to language modelling. In this way, BERT learns contextual relations between words from all of the surroundings (left and right) of a given word.

We splitted the labelled dataset into train and test sets and fine-tuned the pre-trained BERT model on the train test. Moreover, for making the BERT learn semantic clues better in the SERPs, in this work we crawled the newspapers’ contents as the second dataset which is similar to those news articles in the SERPs, labelled with weak supervision and used this dataset to fine-tune the BERT model again. To secondly fine-tune the already fine-tuned BERT model, we used weak supervision labels of the news articles as depicted in Figure 1. As the lower-quality labels were being updated by the existing BERT model, the model was expected to learn linguistic clues such as semantic relations and important keywords which belong to a specific category (conservative, liberal, or neutral) simultaneously. The second question is highly related to the first one, thus we directly used the results of the first question which showed the levels of political bias in the SERPs of Bing and Google, then made a comparison among them by using the statistical analysis described in Section 3.3.

To address the third question, we used the whole corpus in the first dataset to investigate if the input bias exists. To track the source of bias, we made use of the final (fine-tuned twice) BERT model to label all the SERPs in the first dataset and check if the bias in the top 10 documents is consistent with the whole corpus.

---

[2]https://pythonhosted.org/feedparser/
Table 1. Perspective dataset

| query | document title | document content | Weak Supervision Without | Weak Supervision With |
|-------|----------------|------------------|--------------------------|-----------------------|
| ✓     | ✓              | ✓                | 0.647                    | 0.692                 |
| ✓     | ✓              | ✓                | 0.692                    | 0.692                 |
| ✓     | ✓              | ✓                | 0.676                    | 0.707                 |
| ✓     | ✓              | ✓                | 0.707                    | 0.723                 |

4.3 Results

Exp. 1: Fine-tuning BERT on only perspective data using the following inputs:
  a) only document content with title
  b) concatenating query and document content with title
  c) only document title
  d) concatenating query and document title

Exp. 2: Fine-tuning BERT on only news data using the following inputs:
  a) only document content with title
  b) only document title

REFERENCES

[1] [n.d.]. AllSides - Don’t be fooled by media bias and fake news. https://www.https://www.allsides.com/unbiased-balanced-news/. Accessed: 2019-08-20.
[2] [n.d.]. InternetLiveStats. http://www.internetlivestats.com/. Accessed: 2018-10-06.
[3] [n.d.]. Search Engine Statistics 2018. https://www.smartinsights.com/search-engine-marketing/search-engine-statistics/. Accessed: 2018-10-06.
[4] J Shane Culpepper, Fernando Diaz, and Mark D Smucker. 2018. Report from the Third Strategic Workshop on Information Retrieval (SWIRL).(2018), (2018).
[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018).
[6] Alejandro Diaz. 2008. Through the Google goggles: Sociopolitical bias in search engine design. In Web search. Springer, 11–34.
[7] Robert Epstein and Ronald E Robertson. 2015. The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. Proceedings of the National Academy of Sciences 112, 33 (2015), E4512–E4521.
[8] Gizem Gezici, Aldo Lipani, Yucel Saygin, and Emine Yilmaz. 2021. Evaluation metrics for measuring bias in search engine results. Information Retrieval Journal 24, 2 (4 2021), 85–113. https://doi.org/10.1007/s10791-020-09386-w
[9] Eric Goldman. 2008. Search engine bias and the demise of search engine utopianism. In Web Search. Springer, 121–133.
[10] Juhi Kulshrestha, Motahhare Eslami, Johnnatan Messias, Muhammad Bilal Zafar, Saptarshi Ghosh, Krishna P Gummadi, and Karrie Karahalios. 2018. Search bias quantification: investigating political bias in social media and web search. Information Retrieval Journal (2018), 1–40.
[11] Abbe Mowshowitz and Akira Kawaguchi. 2002. Assessing bias in search engines. Information Processing & Management 38, 1 (2002), 141–156.
[12] Safiya Umoja Noble. 2018. Algorithms of Oppression: How search engines reinforce racism. NYU Press.
[13] ProCon.org. 2018. https://www.procon.org
[14] Ryen White. 2013. Beliefs and biases in web search. In Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval. ACM, 3–12.