Bauhaus-Universität Weimar
Faculty of Media
Degree Programme Medieninformatik

Modelling and Evaluating Bias in Search Engines

Bachelor's Thesis

Valerie Lemuth

1. Referee: Prof. Dr. Benno Stein
2. Referee: PD Dr. Andreas Jakoby

Submission date: December 24, 2020

Declaration

Unless otherwise indicated in the text or references, this thesis is entirely the product of my own scholarly work.

Weimar, December 24, 2020

. . . .
Valerie Lemuth

Ich hol gleich meinen Brockhaus raus
Und dann googlen wir
Im Papier —Rainald Grebe

Abstract

Search engines are a part of everyday life and for many they have become indispensable. However, users often don't realize the influence search engines can have on the search results they interact with. To analyze this influence, an approach for assessing bias in search engines is developed in this thesis and then used in an experiment with the search engine Google. The approach focuses on one search engine at a time. It looks at the search results the search engine produces regarding specific topics, during a defined time frame, and at a selected location. To calculate the probability of a bias, often asked queries are selected for each topic, and the search results are compared with multiple references, using statistical methods. The references come from different sources, most important however is that they include the opinion of society on the topic. The approach is successfully used to asses the bias of Google for nine topics. For two out of the nine topics the findings show that a bias is likely.

Contents

1 Introduction **1**

2 Related Work **4**
2.1 About Search Engine Bias 4
2.2 Measuring Bias in Information Retrieval Systems 5

3 Approach for Assessing Bias **10**
3.1 Process 10
3.1.1 Variables, Confounders and Limitations 14
3.2 Mathematical Description 14
3.2.1 Aggregated Stance Calculation 15
3.2.2 Reference Distribution Simulation 16
3.2.3 Bias Assessment 17

4 Experiment and Evaluation **20**
4.1 Selection of Topic, Query and Reference Distribution 20
4.1.1 Topics 20
4.1.2 Queries 21
4.1.3 Reference Distributions 24
4.2 Findings and Discussion 27
4.2.1 Aggregated Stance 27
4.2.2 Simulated Reference Distribution 35
4.2.3 Bias Assessment 35

5 Conclusion and Future Work **41**

A Simulated Reference Distributions **43**

Bibliography **48**

Acknowledgements

First and foremost I would like to thank Prof. Dr. Benno Stein, Yamen Ajjour and Martin Potthast for providing the foundation of this thesis and the ongoing support throughout.

Prof. Dr. Benno Stein and Dr. Andreas Jakoby I would like to thank for examining my work.

Further, I would like to thank Yamen Ajjour for providing the stance classifier used in this thesis.

Many thanks also go to my friends and especially to Johanna Sacher for walking this path alongside me. We did it!

My endless gratitude goes to my mother Dr. Karin Lemuth for proofreading this thesis far too many times and for supporting me in every aspect of my life.

Last but not least I would like to thank Pascal Welp for supporting and loving me, whenever I need it.

Chapter 1

Introduction

In this thesis, an approach for assessing bias in search engines is constructed and then used in an experiment analysing the search engine Google.

The collection and storage of information is nothing new when looking at the history of humanity. However, this access to information has always been limited. With the emergence of the internet, this has changed. Almost everyone has access and everyone can add to this ever-growing system. Search engines bring order to the chaos and help users find what they are searching for. It is widely known that for this purpose different types of algorithms for crawling, indexing and ranking are used. More than this is oftentimes not known to the user, and so for many a search engine is more like a black-box. With the use of algorithms, it is decided which results are presented in response to a query and which are presented at the top of the search engine result page (SERP). This order gives an impression of different importance's, implying that a result with a higher position has greater relevance and is of more importance.

Different researchers have demonstrated that search engines have a lot of influence on users and the society in general. Studies have shown that users have great trust in the ranking of search engines, and only a few seek out the relevant results to their query (Pan et al. [2007], Keane et al. [2008]). People are also influenced by the presentation of the results on the SERP (Novin and Meyers [2017]). As search engines want to keep the user on their platform as long as possible, they try to return those results that will most likely please them (White [2013]). In 2013 Ryen White showed that users prefer positive information over negative and that in turn, search engines return positive results, regardless of the truth. This can then lead to users accepting false information (White [2013]). In 2015 Epstein and Robertson even came to the conclusion that search engines could influence the results of an election (Epstein and Robertson [2015]). Important to note is that the described influence of search results on users does not need to be immediate. It is comparable to the effect of advertisements and relies on a large amount of people which are influenced to a certain extend, rather than one person's opinion changing immediately. With over 6 billion daily searches on Google [1], this process is continuous and happens in parallel.

The approach for assessing bias, described in this thesis, focuses on one search engine at a time. It looks at the search results the search engine produces regarding specific topics, during a time frame, and at a selected location. Only ad-hoc retrievals are taken into account. Further, the focus lies on detecting bias, which has the potential to influence the society at the selected location. The studied search engine should therefore be popular and often used at the location. A bias in search engines would also be especially relevant for controversial topics since these are per definition the subject of intense public disagreement [2]. As described above, search engines could further intensify this disagreement or support a side through the search results they return for queries concerning a topic. The controversy of a topic also depends on the location and can differ from country to country. Controversial topics often have two distinct sides, which also helps during the bias assessment, since each topic can be divided into pro, con, and neutral. Therefore, the question this thesis asks is whether search engines deliver biased search results on controversial topics.

Bias can be seen as a distortion of the truth, and usually, a so-called ground truth is used to measure this distortion; in this thesis, the use of reference distributions is proposed instead. When assessing search engine bias for a topic, the stances of the search results need to be studied. Often it is assumed, that the stances should have an equal distribution for the search results to be unbiased. We propose that the opinion of society or statistics on the topic should be reflected in the stance distribution, accordingly. Therefore, instead of using one ground truth per topic, multiple reference distributions from different sources are used for the bias assessment within this work. For each topic, the reference distributions portrait stance distributions or facts regarding the topic from different sources. To the best of our knowledge, this is a novel approach. We define bias as the deviation from the reference distribution. We define search results as biased, if the stances of the search results deviate from the expected reference distribution and only see this as being relevant, if there is a possible influence on society.

During the bias assessment, the behaviour of the search engine regarding

[1] `https://techjury.net/blog/google-search-statistics/` [accessed September 22, 2020]

[2] `https://www.collinsdictionary.com/de/worterbuch/englisch/controversial` [accessed September 27, 2020]

each controversial topic is studied separately. For each topic, the reference distributions are compared with the stances of the SERPs, produced by the search engine. For this, suitable queries, which are posed to the search engine, first need to be selected. It is important that these are actual queries on the topic, which are posed by real users and that they are often asked. Rarely asked queries result in a SERP that is seldom seen and therefore a possible bias is not as relevant in regards to society. For each of these top queries, the stances of the search results on the SERP are inferred using a stance classifier. The possible stances of a search result are either pro, con or neutral and within this work, they are represented as 1, -1 and 0. Based on the ranking of the search results, an aggregated stance for the whole SERP is calculated. For this, the stances are weighted according to their position on the SERP and then added up. This is because results at higher positions seem more relevant to the user and their stances need to be weighted accordingly.

For each topic, the bias assessment consists of two parts, the comparison of the queries with the reference distributions and the calculation of the bias for the whole topic. As a first step, the aggregated stance of each query is compared with each reference distribution. To make them comparable, each reference distribution is simulated and a distribution of aggregated stances is created. Then, for each reference distribution the probability of the aggregated stance of each query belonging to the simulated reference distribution is calculated. To asses the bias of the search engine for the whole topic, these probabilities are weighted according to the importance of the query and then added up. This results in the probability that the ranking of the search results, seen by an arbitrary user of the search engine, on a given topic, at the time and location of measuring, reflects the expected reference distribution. The higher this probability, the more likely that there is no bias, in the search engine, for the topic, at the time and location of measuring.

It is important to note, that multiple variables exist, which can be somewhat controlled, but still influence the outcome of the presented approach. For example, a user's location, language or search behaviour. As these differ for every user, it is impossible to truly get an accurate idea of the bias each user experiences. This can only be simulated to a certain degree. The stance classifier and its error rate will also inevitably be reflected in the results. This is because of technical limitations and can be improved in future works.

Within this thesis, an experiment is conducted, using the described approach, while concentrating on the US and Google between 2019-2020 (see chapter 4). For this, nine controversial topics with relevance in the US are chosen. The reference distributions are constructed of the categories *Opinion Poll*, *Market Information*, and *Political Landscape*. The bias assessment for the nine topics is presented and an outlook is given.

Chapter 2

Related Work

Bias, especially in search engines, has been widely discussed over the last years. Papers, ranging from its inevitability to methods for detecting it are presented within this chapter.

2.1 About Search Engine Bias

Some researchers argue that search engine bias is inevitable, necessary or even desirable. Dirk Lewandowski, for example, discusses the responsibilities of Google in regards to bias in search results. He comes to the conclusion that because of the sheer size of the web, "we cannot expect a search engine to provide fair and unbiased results" (Lewandowski [2017] : 16). Through crawling and indexing, a bias naturally happens, this is especially apparent when comparing search engines, since they use different algorithms and therefore have different results for similar queries (Lewandowski [2017]).

A similar stance is taken by Karsten Weber. He argues that user behaviour, indexing, and the content of the web itself, all lead to a bias in web search engines. Like Lewandowski, he advertises the use of multiple search engines and not exercising blind trust (Weber [2011]).

Lewandowski and Weber both shine a light on *unintentional* reasons for distorted search results, yet this way, they disregard the possibility of an intentional bias and the further implications this holds.

Edelman shines a light on another aspect that could lead to a bias in search results. Search engines are businesses with interests that they want to support as much as possible. He focuses on Google and explains that as a company they have reasons to build traffic to the services they provide and to protect their advertising system (Edelman [2011]).

Eric Goldman even goes so far as to say that search engine bias is desirable. In his essay, he argues that search engines need to make editorial choices in the SERP to filter out spam and to optimize their search results for the users, otherwise they would simply move on to a "better" search engine. This leads to a necessary bias. The bias that he is referring to is the effect of big websites getting bigger and small websites staying small. In his opinion, most bias will be disappearing with the introduction of personalized ranking algorithms since different users will be receiving different search results for the same queries (Goldman [2006]). Important to note, is that this approach ignores the effect of filter bubbles and even idealizes them. Search engines that concentrate on only returning those results which will most likely please the user, create filtered bubbles of opinions that fit with the users' ideals. Most of the time, users are not aware, that the results they are receiving are biased. Personalized search engines, as explained by Goldman, also lead to what some researchers call user bias, meaning users getting different results based on gender, race, and so on (Pitoura et al. [2018]).

2.2 Measuring Bias in Information Retrieval Systems

With many different definitions of bias, multiple papers have been written on the detection of bias. Many researchers use different terminologies or information retrieval systems while talking about the same concepts. In the effort of giving an overview of different papers and explaining the terms in the context of this thesis, table 2.1 is provided.

Mowshowitz and Kawaguchi have been looking at search engine bias as early as 2002. They created a system for assessing bias in search engines using URLs. Their definition of bias calls for a norm (a type of reference distribution) to act as a baseline for the bias detection. They used a set of search engines to create a distribution of URLs for the results of each query. This acts as a reference distribution. A specific search engine can then be compared to this and the bias can be inferred. Overall they looked at thirteen search engines and were able to find differences in the resulting URL distributions of the different search engines (Mowshowitz and Kawaguchi [2002]). Although this approach is good for comparing search engines, and the result-diversity, it doesn't detect bias as defined in this thesis. This thesis looks at the stances present in the results and compares them to a reference distribution. Mowshowitz and Kawaguchi show whether a search engine is able to find the same results for specific queries as an other set of search engines. If this is the case, it would then mean that the search engine isn't biased. In our opinion, bias assessment needs to be based on multiple different reference distributions, where the opinion of society is of big importance, since this is where the bias has an impact.

Chen and Yang took a similar approach. They based their work on Mowshowitz and Kawaguchi but defined bias in search engines as consisting out of two parts, the indexical and the content bias. They defined the indexical bias as the bias described by Mowshowitz et al. and used the same method for detecting it. For the content bias, they introduced a new method based on weighted vectors that represent the content of the web pages. Using HTML tags of websites, they created query-specific representative vocabulary vectors (RVV) for the different search engines. Like Mowshowitz et al., they created a reference distribution using comparable search engines, they used the RVV of the websites for the content bias. While looking at ten randomly chosen *hot* queries from Lycos, they were able to show a statistical significance of both, indexical and content bias (Chen and Yang [2006]). Chen and Yang detect bias two ways, using URLs and HTML tags. They could be seen as two reference distributions, similar to the approach described in this thesis. However, their reference distributions are based on the behaviour of a set of search engines compared to the one search engine being investigated. This again compares search engines, the type of results and content they return for queries, but this is not a bias assessment, as defined in this thesis.

Kulshrestha et al., on the other hand, looked at the Twitter search engine in the context of the 2016 US presidential elections. This results in two possible stances, republican and democratic. To detect the search engine bias, they made a difference between the input, output and ranking bias. The input bias, they defined as the bias of all data relevant to a query, the output bias, as the bias in the data collection output by the search engine for a query, and the ranking bias, as the bias solely produced by the ranking system. In the context of this thesis, the input bias would be called the aggregated stance of all data and the output bias the aggregated stance of the collection retrieved by the search engine. The ranking bias is the only one that is comparable to our definition of bias, meaning a deviation from a reference distribution. Since they calculate the ranking bias as the difference between the output and the input bias, the input bias could be called a reference distribution. To detect the stance of a tweet, they created sets of known republican and democratic Twitter users and compared the interests of the author of said tweet with the interests of the two sets using weighted vectors. For 25 republican and democratic queries, asked repeatedly over the course of one week, they were able to infer that the ranking system shifts the stance of the results towards the party of the query. This shows a bias in the ranking system since the aggregated stance of all data was generally more democratic leaning. For some republican queries, the ranking system even shifted the aggregated stance of the results from democratic to republican (Kulshrestha et al. [2017]).

Based on the work described above, Kulshrestha et al. released an extended version of the paper two years later, including an approach for detecting bias in web search engines, in this case specifically, Google (Kulshrestha et al. [2019]). For this, they used the same topic and queries as before. As they don't have access to the input data of web search engines, they only concentrated on the output bias. In the context of this thesis, this would be called the aggregated stance of the SERP. To infer these stances they used existing lists which already categorized the stances of political news media websites. Wikipedia articles they counted as neutral and websites of politicians as the corresponding party. It could be seen that Google results lean towards the party of the query (Kulshrestha et al. [2019]). As mentioned above, the bias that Kulshrestha et al. measure, is what we define as the aggregated stance. This is only our first step for detecting bias, since bias is relative to something and needs to be based on a ground truth or reference distribution.

Pitoura et al. offer a theoretical approach for detecting bias in online information providers with multiple options for different steps. Similar to the method described in this thesis, they base the computation of bias on a ground truth and propose a type of machine-learning for detecting the stances of the results. Yet their approach also includes methods for detecting a user bias, meaning a bias against a person because of their gender, race, and so on. The possible topics they talk about, need to have differentiating attributes and the resulting queries simply need to be based on them, using a *knowledge base* (Pitoura et al. [2018]).

Table 2.1: Paper about measuring bias in information retrieval systems in the context of this thesis

	This Thesis	**Mowshowitz and Kawaguchi [2002]**	**Chen and Yang [2006]**
Information Retrieval System	Web search engines (Google)	Thirteen comparable web search engines	Ten web search engines
Bias Definition	Deviation of stance distribution from reference distribution	Balance of items in collection retrieved from database	**Indexical Bias:** Mowshowitz et al. **Content Bias:** Difference of page content
Reference Distribution	Multiple, from different sources, with the opinion of the population as an important part	Distribution of URLs of set of comparable search engines for queries	**Indexical Bias:** Mowshowitz et al. **Content Bias:** Distribution of weighted vectors (RVV) representing the web pages
Topics	Controversial topics	4 topics	-
Query Selection	Most asked queries for each topic	12 queries for each topic	Ten random *hot* queries from Lycos 50
Stance Detection	Stance classification for each search result & aggregated stance calculation for each query	Getting URL collection for each query	**Indexical Bias:** Mowshowitz et al. **Content Bias:** Create RVV based on Html-tags
Bias Detection	Compare aggregated stance with reference distribution	Deviation of URL collection from reference distribution	**Indexical Bias:** Mowshowitz et al. **Content Bias:** Similarity between RVV and reference distribution

Table 2.1: Paper about measuring bias in information retrieval systems in the context of this thesis (continued)

	Kulshrestha et al. [2017]	**Kulshrestha et al. [2019]**	**Pitoura et al. [2018]**
Information Retrieval System	Twitter search engine	Web search engines (Google)	Online Information Provider (OIP)
Bias Definition	**Input Bias:** Aggregated stance of data input to ranking system **Output Bias:** Aggregated stance of collection produced by ranking system **Ranking Bias:** Difference between output & input bias	**Output Bias:** Aggregated stance of SERP for query	**User Bias:** Different users receiving different information based on race, gender, etc. **Content Bias:** Bias in data received by the user
Reference Distribution	Input bias		- Compare all data on topic with data selected by OIP - Crowd sourcing - Estimation of stance distribution in population
Topics	2016 US election Republicans vs Democrats	2016 US election Republicans vs Democrats	Any topic with different sides
Query Selection	25 hashtags	25 hashtags	**Query Generator:** Create queries for topic and different stances using knowledge base
Stance Detection	Comparing author of tweet to known set of republican and democratic users	News media website = check in existing list with stances, Wikipedia = neutral, Website of candidate = party of candidate	**Result Processing:** Determine stance of results using machine-learning
Bias Detection	Calculate ranking bias (Average over one week)	Aggregated stance of SERP (Average over one week)	Compute bias using bias metrics and reference distribution

Chapter 3

Approach for Assessing Bias

In this thesis, an approach for assessing bias in search engines was developed. Within this chapter, the approach is described, starting with the whole process to gain an overall view, followed by the mathematical description (aggregated stance calculation, reference distribution simulation and bias assessment).

3.1 Process

In Figure 3.1 an overview of the process is depicted, it starts with the selection of controversial topics, for which the top queries are inferred and then posed to the chosen search engine. The stances of the search results are derived using a stance classifier. Based on this, the aggregated stance for each query is calculated. In parallel, for each topic, a reference distribution is created and simulated to create a distribution, comparable to the aggregated stances. Finally, the probability of each aggregated stance belonging to the simulated distribution is calculated. With the summation of the probabilities for each topic, the possibility of a bias can be assessed. It is important to note, that the here described process for assessing bias in search engines is based on a time frame and a location. Depending on this, a popular and often used search engine is selected. This is because the focus of this thesis lies on the assessment of bias, which has the potential to influence society. During the process, it is looked at where society could be impacted the most by a bias and then the next step is chosen accordingly. For example, the topics are picked based on the country, their importance, controversy and impact. Further, the process described in this thesis only assesses the bias of one search engine at a certain location and time frame, for the chosen topics.

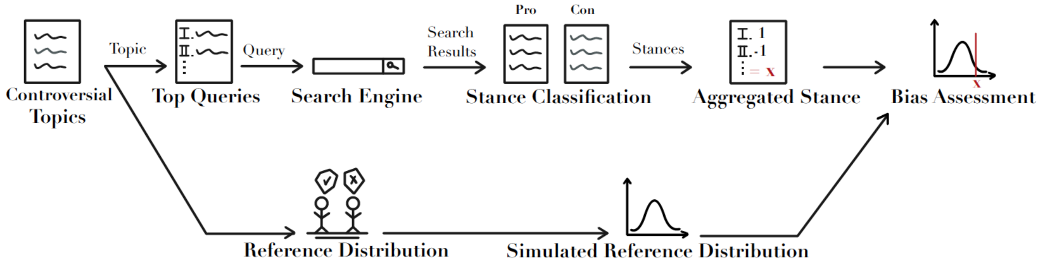

Figure 3.1: Process Overview

In the following, important steps of the process are described in more detail.

Controversial Topics

At the beginning of the process, controversial topics are chosen. Depending on the location, selected for the bias assessment, the topics should be current, since, for outdated topics, a possible bias will not matter as much in regards to society. But more importantly, they should be controversial, because such topics are already heavily debated, often with two distinct sides. Controversial topics often divide society and a bias in a prominent search engine could further intensify this gap. For each topic there ideally should be a clear pro and con side. If this is not possible, the opinions should be grouped into general pro and con sides.

Top Queries

For each topic, the most asked queries are selected. This again depends on the search engine and the location. It is also important that these are actual user queries because only such queries guarantee an accurate bias assessment. The SERP of rarely asked queries is seldom seen and therefore they have minimal influence on society, that is why they should not be used. Depending on the search engine, there may exist platforms which offer insight into the most asked queries per topic, at a selected location. A finite amount of these top queries needs to be selected and then weighted according to how often they are asked (see section 3.2.1).

Aggregated Stance

Each query is then posed to the search engine, the stances of the search results are detected and based on them, the aggregated stance, for a query, is calculated. A detailed description of the formulas used within this step can be seen in section 3.2.1. The stances, of a constant amount K of search results, are inferred using a stance classifier. The possible stances are pro, con and neutral and in this thesis, they are represented by the numbers 1, -1 and 0. After each search results is assigned its stance, the aggregated stance for a query is calculated. For this, the stances are weighted according to their position on the SERP and then summed up (see chapter 3.2.1). In this approach, the stances of search results at higher positions weigh more than those of the results at the bottom of the page. This is because results at higher positions seem more relevant to users and their content seems *more correct*. This needs to be reflected in the bias assessment. For each query of a topic, the aggregated stance roughly shows the overall stance of the SERP, by its sign. This aggregated stance is then used for the bias assessment.

Reference Distribution

Independent from the query selection and the aggregated stance calculation, multiple reference distributions are created for each topic. For a topic, each reference distribution comes from different sources. A reference distribution could be a stance distribution or facts on the topic. If possible, each reference distribution should be transferred into an appropriate pro, con and neutral scheme in percent. Within this thesis, these reference distributions are used for the bias assessment, instead of a ground truth. This is because we argue that there is not one specific truth when working with search engines. A bias in the search results, seen by users, would have an impact on the opinion of society. Therefore, we see it as important to include society's opinion as a part of the reference distribution. Further, other sources can be used, for example, statistics or facts about the topic or political decisions. The aggregated stances are based on the search results returned by a search engine and the reference distributions come from external sources. Therefore, the reference distributions act as a baseline, and the aggregated stances can be compared to it. Depending on the source, some reference distributions are better applicable for the topic than others.

Simulated Reference Distribution

To be able to compare the aggregated stance of a query with the reference distributions, each reference distribution needs to be simulated based on the same formulas used for the aggregated stance calculation (for more detail, see section 3.2.2). Each reference distribution is simulated, and used for the bias assessment, separately. For this, an artificial SERP is created, with a constant amount K of search results. This amount is the same constant amount K, used during the aggregated stance calculation. For each of the K simulated search results, the stance is *guessed* based on the reference distribution. Each simulated search result gets assigned a stance, based on the pro, con and neutral percentage distribution. The aggregated stance is then calculated for the simulated SERP. This is repeated a thousand times, to create a distribution of aggregated stances and their frequencies. The amount of repetitions is variable, but it should be relatively large, to create a good distribution of values.

Bias Assessment

For a more detailed, mathematical description of the bias assessment, see section 3.2.3.

Probability Calculation for each Query

For each reference distribution, the probability of the aggregated stance of each query belonging to the simulated distribution is calculated (see section 3.2.3). For each topic, exist multiple reference distributions, which have all been simulated to create a distribution of aggregated stances. For each topic, there also exist multiple top queries for which the aggregated stance has been calculated as well. For each query, the probability of the aggregated stance belonging to each simulated distribution is calculated. This probability is calculated using a permutation test (for more details, see section 3.2.3).

Probability Calculation for each Topic

For each reference distribution, the probabilities of all top queries, belonging to a topic, are weighted according to the importance of the query and then summed up (see section 3.2.3). This results in the probability of the top queries belonging to a reference distribution. With the top queries, we assume to cover 90% of all user inquires since we stop with the query selection after a finite amount. The other 10%, for which we don't calculate the bias, we assume to be to the benefit of the search engine and use the probability of the expected value. In the end, there are the probabilities of a topic belonging to each reference distribution. Based on those results, a conclusion about a bias can be drawn.

3.1.1 Variables, Confounders and Limitations

Variables

Since the user's location and language influence the search results, these variables should be controlled as best as possible. E.g. by posing the queries always using the same location and corresponding language. The location also needs to be considered when selecting suitable topics for the bias assessment. A topic like gun control is very controversial in the US, yet hardly so in Germany. The search engine should also be selected with the location in mind. Google is very popular in America but when looking at China, the search engine Baidu is the market leader[1].

Confounders

In this thesis, the closeness to reality in every step is very important, to create the best possible bias assessment. But this can only be simulated to a certain degree, as for each user, the user behaviour and specific location lead to different search results. When looking at a whole country, individual user behaviour, e.g. dwell time or which results are viewed, is impossible to simulate. Different cities in the same country may also result in different SERPs and therefore this also acts as a confounder. The stance classification is also error-prone, as it can only be as good as the state of the art. The error rate needs to be taken into account and the results of the classifier need to be spot-checked to get an impression of the distortion.

Limitations

This approach is limited to detecting the bias of a search engine for specific topics. It is not possible to categorize the whole search engine as either biased or unbiased.

3.2 Mathematical Description

In this section, the mathematical description of the process is presented. To improve the legibility, a table of notations is provided in the following (see table 3.1).

[1] `https://komarketing.com/blog/top-3-chinese-search-engines/` [accessed December 2, 2020]

Table 3.1: Notation

Symbol	Meaning
$t \epsilon T$	Topics
$q \epsilon Q_t$	Queries per topic
$f(q)$	Query frequency
$w(q)$	Query weight
D	All documents indexed by the search engines
Q	All queries submitted to the search engine
$\rho : Q\, x\, D \rightarrow [1, \|D\|]$	Ranking function of the search engine
$\rho(q, d) = i$	Rank of document under query
d_i	Document at rank i under query
$d \epsilon D_q$	Documents per query
$\sigma : D\, x\, T \rightarrow \{-1, 0, 1\}$	Stance function
$\sigma(d, t)$	Stance of document for topic
$as(q)$	Aggregated stance for a query
$r \epsilon R_t$	Reference distributions per topic
P_r	Simulated reference distribution
S_r	Sample of simulated reference distribution
$\overline{s}$	Mean of sample S_r
$p_r(as(q))$	Probability that aggregated stance belongs to P_r
$p_r(t)$	Probability that topic belongs to P_r

3.2.1 Aggregated Stance Calculation

In this section, the weighting of the queries and the calculation of the aggregated stances are described in detail. The queries are weighted according to how often they are asked and the search results for each query are weighted according to their position on the SERP, and then summed up. The weighted search results, result in the aggregated stance of each query.

Query Weight

For each topic $t \epsilon T$, the top queries $q \epsilon Q_t$ are weighted according to how often they are asked, with the most asked query getting the highest query weight $w(q)$. Depending on the used search engine and platform, these weights could already be given or could still need to be converted into the needed format.

Aggregated Stance Calculation

For each query $q \epsilon Q_t$, the aggregated stance as(q) is calculated by weighting the stances of the search result documents $d \epsilon D_q$, using the *Discounted Cumulative Gain (DCG)* [2]. The DCG_k is the total gain at the rank position k, with

[2] https://dl.acm.org/doi/10.1145/582415.582418

the relevance of a document at rank i rel_i and the reduction factor $log_2(i+1)$:

$DCG_k = \sum_{i=1}^{k} \frac{rel_i}{log_2(i+1)}$

For the aggregated stance, rel_i is defined as the stance $\sigma(d_i, t)$ of the search result document d_i at the position i for the topic t. To display the stances pro, con or neutral, $\sigma(d_i, t)$ can either take the values -1, 0 or 1. The rank position k is defined as the constant amount of search results K, used for each query.

Aggregated stance: $as(q) = DCG_K = \sum_{i=1}^{K} \frac{\sigma(d_i,t)}{log_2(i+1)}$; $\sigma(d_i, t) \epsilon \{-1, 0, 1\}$, K = Constant amount of search results

At first glance, the aggregated stance $as(q)$ gives an idea of the overall stance of the investigated SERP. A positive $as(q)$ shows the SERP leaning towards pro and a negative, towards con. The closer the value is to zero, the more neutral results are present in the SERP. In case the $as(q)$ is zero, this means, that either all the results are neutral or the pro and con results are arranged exactly in such a way that they are well balanced and mathematically cancel each other out. Either way, if this is the case, the SERP doesn't favour one side over the other. In the following, the aggregated stance is illustrated by an example.

Example using random stances:

$\sigma(d_i, t)$: -1, -1, 1, 0, -1, 0, 1, -1, -1, 1

DCG_i : -1, -1.6, -1.1, -1.1, -1.5, -1.5, -1.2, -1.5, -1.8, -1.5

$as(q) = DCG_{10} = -1.5$

Using random stances, with a higher amount of con results at higher positions, it can be seen that the $as(q)$ results in a negative number and correctly portraits this overall stance.

3.2.2 Reference Distribution Simulation

For the bias assessment, the aggregated stances and the reference distributions are compared; to achieve this, each reference distribution is simulated. For each topic $t \epsilon T$, there are multiple reference distributions $r \epsilon R_t$. As described above, a reference distribution is composed of a pro-, con- and neutral-distribution in percent. Based on this percentage distribution, the stances of K artificial search results, on a simulated SERP, are randomly chosen. K equals the constant amount of search results used in section 3.2.1. With the K simulated stances, the aggregated stance, as described in section 3.2.1, is calculated. For each $r \epsilon R_t$ this is repeated ≥ 1000 times to create a simulated distribution P_r of aggregated stances and their frequencies. The sample S_r of the simulated reference distribution P_r, consisting only of the aggregated stances, is used in further sections.

Example using a $[0.60, 0.30, 0.10]$ distribution:

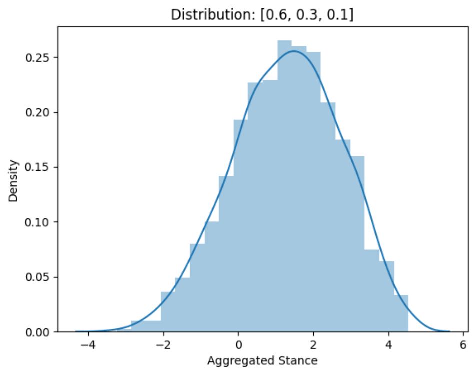

Figure 3.2: Graph of simulated distribution $P_{[0.60,0.30,0.10]}$, x-axis: aggregated stances, y-axis: estimated kernel density

The graph in Figure 3.2 displays the simulated distribution $P_{[0.60,0.30,0.10]}$ based on a reference distribution with 60% Pro, 30% Con and 10% Neutral. The reference distribution was simulated using $K = 10$ search results and repeating the process a 1000 times. The aggregated stances are displayed on the x-axis and the kernel density, the estimated probability of each aggregated stance, on the y-axis. As can be seen, the range of aggregated stances goes from about -3 to about 5, meaning the majority of aggregated stances are positive. Since 60% of the reference distribution is pro, the graph correctly reflects this.

3.2.3 Bias Assessment

In this section, the bias assessment of a search engine, for a topic, at a selected location and time is described. For each reference distribution $r \epsilon R_t$ of a topic t, the probability $p_r(as(q))$ is calculated for all queries $q \epsilon Q_t$ on the topic. $p_r(as(q))$ represents the probability of the aggregated stance $as(q)$ belonging to the simulated reference distribution P_r. Then the probabilities $p_r(as(q))$ of all queries $q \epsilon Q_t$ belonging to a topic are weighted according to the corresponding query weight $w(q)$ and then summed up. For each topic $t \epsilon T$, this results in the probability $p_r(t)$, which can then be used for the bias assessment.

Probability Calculation for each Query

For each topic $t \epsilon T$, the probability $p_r(as(q))$ of the aggregated stance $as(q)$ of each query $q \epsilon Q_t$, belonging to the simulated reference distribution P_r is calculated. This is done by permuting the sample S_r, n times and each time comparing a single sample $s \epsilon S_r$ with the aggregated stance $as(q)$ using the mean $\overline{s}$ of S_r. For this, s is subtracted from $\overline{s}$ and it is checked whether it's absolute is bigger than the absolute of $\overline{s} - as(q)$. Then, it is counted how often this is the case, for the n permutations, and this number is divided by n to get the probability $p_r(as(q))$.

In the following this is displayed in pseudo-code in the effort of giving a more precise description:

Algorithm 1 Pseudo-Code

for t in T **do**
    for r in R_t **do**
        for q in Q_t **do**
            $n = 1000$
            for $i = 1, 2, \ldots, n$ **do**
                $permute(S_r)$
                if $|\overline{s} - s| > |\overline{s} - as(q)|$ **then**
                    $count \leftarrow count + 1$
                end if
            end for
        end for
        $p_r(as(q)) \leftarrow \frac{count}{n}$
    end for
end for

Probability Calculation for each Topic

To get the overall probability $p_r(t)$ of a topic t belonging to a simulated reference distribution P_r, the probability $p_r(q)$ of all top queries $q \epsilon Q_t$ are weighted according to the corresponding query weight $w(q)$ and then summed up. With the top queries we assume to cover about 90% of all user inquiries, the 10% we don't cover, we assume to be to the benefit of the search engine. Therefore we assume that the expected value of the distribution P_r, the mean, is met for those 10% and therefore calculate a probability of 1.0.

Probability $p_r(t) = 0.9 * (\sum_{q \epsilon Q_t} w(q) * p_r(q)) + 0.1 * 1.0$

In the context of the bias assessment, $p_r(t)$ is the probability, that the ranking of the search results, seen by an arbitrary search engine user, on a given topic t, in a time frame and at a location, reflects the expected distribution $r \epsilon R_t$. The higher $p_r(t)$, the higher the probability that there is no bias.

Chapter 4

Experiment and Evaluation

In this chapter, the described approach is used, while focusing on the US, between 2019-2020 and Google as the search engine. In the first section, the selection of the topics, queries and reference distributions is described. The second section shows the calculation of the aggregated stance for each query, using a stance classifier. After this, follow the reference distribution simulation and the bias assessment.

4.1 Selection of Topic, Query and Reference Distribution

For the experiment, nine controversial topics were chosen, the most asked queries were identified and the reference distributions were selected. The topics with the most asked queries can be seen in table 4.1 and 4.1. The reference distributions are portrayed in table 4.2.

4.1.1 Topics

Using Wikipedia and Statista, nine controversial topics, relevant for the society of the United States, were chosen. In the list of controversial issues [1] by Wikipedia, topics are listed, for which the corresponding articles are edited often and which, therefore, are rated as controversial. In combination with statistics presented by Statista [2], important topics, for which there is a notable difference of opinion in the US, were chosen. They can be seen in table 4.1 and 4.1.

[1] `https://en.wikipedia.org/wiki/Wikipedia:List_of_controversial_issues` [accessed November 20, 2020]

[2] `https://de.statista.com/themen/117/usa/` [accessed November 20, 2020]

4.1.2 Queries

The platform Google Trends[3] was used, to select the most asked queries for each topic, then the query weights were calculated, based on the frequency given by Google Trends. For each topic, the queries, the frequencies and the corresponding weights can be seen in table 4.1. With Google Trends, the search behaviour of specific countries and the trend of search terms at selected locations and time frames can be studied. When looking at the trend of a search term, Google Trends offers two options, searching as a term or a topic. A term, returns the trend for all search terms that match it, a topic, groups together all search terms that belong to it, without them needing the same wording or even language[4]. In both cases, a graphical overview of the trend, at the selected location and time frame is presented. Related queries and their frequency can be seen as well. Whenever possible, the search for the topics was done as a Google Trends topic. This way the related queries concern the whole topic and are not just variations of the search term. Only for the topic *Mandatory Vaccinations* this was not possible. During the search, the location was set to the US and the time frame to the last 12 months. So, for each of the nine topics the most asked queries were selected. For each related query, Google Trends gives the frequency of how often it is asked: A query with the frequency 100 is asked the most and a query with the frequency 50 is asked half as often. We selected the most asked queries and stopped when the frequency fell below 10. That is why, the amount of queries, selected for the topics, ranges from three to nine, depending on the frequency.

For each of the nine topics $t \epsilon T$ and the most asked queries per topic $q \epsilon Q_t$, the query weight $w(q)$ was calculated, using the query frequency $f(q)$ given by Google Trends:

$$w(q) = \frac{f(q)}{\sum_{q \epsilon Q_t} f(q)} \text{ with } \sum_{q \epsilon Q_t} w(q) = 1$$

[3] `https://trends.google.de/trends/?geo=US`

[4] `https://support.google.com/trends/answer/4359550?hl=en` [accessed November 23, 2020]

Table 4.1: Topics with the corresponding most asked queries, their frequency according to Google Trends and their calculated weights

Topic t	**Most asked queries** $q \in Q_t$	**Frequencies** $f(q)$	**Weights** $w(q)$
Abortion	abortion	100	0.79
	abortions	17	0.13
	pill abortion	10	0.08
Climate Change	climate	100	0.30
	climate change	93	0.28
	warming	47	0.14
	global warming	42	0.12
	climate change is	20	0.06
	global warming is	12	0.04
	what is climate	11	0.03
	global climate	10	0.03
Death Penalty	death penalty	100	0.28
	execution	75	0.22
	the death penalty	44	0.12
	executed	42	0.12
	death row	35	0.10
	executions	19	0.05
	capital punishment	17	0.05
	death penalty states	12	0.03
	death sentence	11	0.03
Gun Control	gun	100	0.49
	gun laws	56	0.27
	gun control	48	0.24
Mandatory Vaccination	is vaccination mandatory	100	0.21
	mandatory vaccination bill	92	0.19
	mandatory vaccination 2020	83	0.17
	mandatory vaccination oklahoma	59	0.12
	mandatory vaccination laws	46	0.10
	stop mandatory vaccination	46	0.10
	mandatory vaccination states	40	0.08
	mandatory vaccination bill 2020	16	0.03

Table 4.1: Topics with the corresponding most asked queries, their frequency according to Google Trends and their calculated weights (continued)

Topic t	**Most asked queries** $q \in Q_t$	**Frequencies** $f(q)$	**Weights** $w(q)$
Marijuana Legalization	marijuana	100	0.34
	legal marijuana	94	0.31
	is marijuana legal	69	0.23
	states legal marijuana	20	0.07
	legal cannabis	14	0.05
Nuclear Power	nuclear	100	0.63
	nuclear power	19	0.13
	energy	15	0.09
	nuclear energy	15	0.09
	nuclear plant	10	0.06
Same Sex Marriage	gay marriage	100	0.56
	gay marriage legal	22	0.12
	gay marriage legalized	13	0.07
	is gay marriage legal	13	0.07
	supreme court	11	0.06
	gay marriage us	11	0.06
	gay wedding	10	0.06
Universal Health Care	uhc	100	0.48
	uhc provider	26	0.12
	uhc login	22	0.11
	universal health	22	0.11
	universal healthcare	20	0.09
	universal health care	19	0.09

4.1.3 Reference Distributions

For each topic t, two to three reference distributions have been constructed $r \epsilon R_t$ from the categories *Opinion Poll*(r_1), *Market Information*(r_2) and *Political Landscape*(r_3) (see table 4.2). These reference distributions were then transferred into a pro, con and neutral scheme, to be used during the bias assessment (see table 4.3).

The category *Opinion Poll* represents stance distributions of the US population on specific topics. For eight out of the nine topics, the opinion polls were easily transferable to a pro, con and neutral scheme. The topic *Climate Change* did not fit as easily into this scheme, since the controversy here is not about supporting or opposing climate change but rather about whether it exists. The opinion "doesn't exist", was classified as pro and "exists" as con, because climate change is a very negative topic for those who believe it exists. People who do not believe it exists, don't see it as negative and therefore their opinion was classified as pro.

The category *Market Information* portraits facts and statistics, regarding the topic and the US. Therefore, these reference distributions are often not correctly transferable to the pro, con and neutral scheme of the topics, as they are based on opinions. For the sake of the experiment, the facts and statistics were still mapped to this scheme, knowing full well that this is not entirely correct. For example, the amount of people who "own a gun", is not directly equal to the people against gun control and the people who "don't own a gun" are not necessarily pro gun control. For the other reference distributions, there are similar miss-matches. But the category "market information" still offers another point of reference and these statistics are nevertheless true and should not be ignored.

The category *Political Landscape* looks at the party distribution of the US Senate in the year 2019. Because of the two party system in the US and since they are, most of the time, of opposite opinions, this is a good match. As the US senate is elected directly by the population[5], it is a good representation of the populations opinion. In 2019, there are 100 members in the Senate, 53 Republican, 45 Democrats and 2 Independent. To create the stance distribution, the general opinion of the republican and democratic party on the topics was used. For the two independent members, the opinions were inferred using Votesmart[6] and then added to the corresponding side. The party distribution of the Senate was directly transferable to the used scheme.

[5] https://www.senate.gov/artandhistory/history/common/briefing/Direct_Election_Senators.htm

[6] https://justfacts.votesmart.org

Table 4.2: Topics with the corresponding reference distributions, sorted by the categories *Opinion Poll*, *Market Information* and *Political Landscape*

Topic t	$r_1 \in R_t$ **(Opinion Poll)**	$r_2 \in R_t$ **(Market Information)**	$r_3 \in R_t$ **(Political Landscape)**
Abortion	46% Pro-Choice 49% Pro-Life 5% Unknown Saad [2019]	-	47% Pro-Choice 53% Pro-Life Diffen [2020], Votesmart [2020], Votesmart [2018]
Climate Change	79% Exists 7% Doesn't exist 14% Not sure YouGov and TheEconomist [2019]	-	47% Exists 53% Doesn't exist Salant [2019], Votesmart [2020], Votesmart [2018]
Death Penalty	60% Acceptable 35% Wrong 5% Depends Gallup [2019a]	53290 Prisoners serving life sentences 2814 Prisoners under sentences of death Davis and Snell [2018]	-
Gun Control	66% Support 26% Oppose 8% No opinion Consult and Politico [2019]	30% Own gun 68% Don't own gun 2% Unknown Parker et al. [2017]	47% Support 53% Oppose Diffen [2020], Votesmart [2020], Votesmart [2018]
Mandatory Vaccinations	72% Pro 19% Con 9% Don't know Murad [2019]	1.1% Unvaccinated children Hill et al. [2017]	-
Marijuana Legalization	66% Support 33% Oppose 1% No opinion Gallup [2019c]	52% Have tried marijuana 48% Haven't tried marijuana MaristPoll [2017]	-
Nuclear Power	49% Pro 49% Con 2% Unknown Reinhart [2019]	36.7% Petroleum 32% Natural gas 11.3% Coal 11.5% Renewable energy 8.5% Nuclear power Administration [2019]	-
Same Sex Marriage	61% Pro 31% Con 8% Unknown PewResearchCenter [2019]	-	46% Pro 53% Con 1% Unknown Diffen [2020], Votesmart [2020]
Universal Health Care	54% Pro 45% Con 1% Unknown Gallup [2019b]	-	47% Pro 53% Con Diffen [2020], Votesmart [2020], Votesmart [2018]

Table 4.3: Topics with the corresponding reference distributions transferred into the pro, con, an neutral scheme, sorted by the categories *Opinion Poll*, *Market Information* and *Political Landscape*.

Topic t	$r_1 \epsilon R_t$ **(Opinion Poll)**	$r_2 \epsilon R_t$ **(Market Information)**	$r_3 \epsilon R_t$ **(Political Landscape)**
Abortion	46% Pro 49% Con 5% Neutral	-	47% Pro 53% Con 0% Neutral
Climate Change	7% Pro 79% Con 14% Neutral	-	53% Pro 47% Con 0% Neutral
Death Penalty	60% Pro 35% Con 5% Neutral	5% Pro 95% Con 0% Neutral	-
Gun Control	66% Pro 26% Con 8% Neutral	68% Pro 30% Con 2% Neutral	47% Pro 53% Con 0% Neutral
Mandatory Vaccinations	72% Pro 19% Con 9% Neutral	98.9% Pro 1.1% Con 0% Neutral	-
Marijuana Legalization	66% Pro 33% Con 1% Neutral	52% Pro 48% Con 0% Neutral	-
Nuclear Power	49% Pro 49% Con 2% Neutral	8.5% Pro 91.5% Con 0% Neutral	-
Same Sex Marriage	61% Pro 31% Con 8% Neutral	-	46% Pro 53% Con 1% Neutral
Universal Health Care	54% Pro 45% Con 1% Neutral	-	47% Pro 53% Con 0% Neutral

4.2 Findings and Discussion

4.2.1 Aggregated Stance

To determine the aggregated stance of each query, the queries were posed to the search engine, the content of the result documents was extracted and a stance classifier was used to infer the stances. Then, based one the stances and the individual positions of the results on the SERP, the aggregated stance was calculated.

Preparation

As explained in section 3.1, the location and the influence of user behaviour needs to be controlled, when posing the queries to the search engine. To circumvent the results being adapted to the user behaviour or any user information, the search engine Startpage.com[7] was used instead of Google, to pose the queries. This was possible because Startpage uses the search results of Google, but does not save any user information and therefore doesn't adapt the results to user behaviour[8]. Startpage also offers the option to choose the search language and filter the search results based on a region[9]. For the experiment, the search language "English" and region "US", was used. Only little information about the region filter is offered by Startpage, therefore, a VPN set to New York City was used as well, when posing the queries. In the settings of Startpage, the amount of search results displayed on one page can be selected. For the experiment, 10 results per page were selected and only the first page was considered for the bias assessment. These 10 results are equivalent to Googles first page and since only few users go to the second page to find what they are searching for, this covers the most seen search results (Baeza-Yates et al. [2005]). We also argue that the first 10 results cover the most influential search results since the relevance of search result content, as seen by users, declines quickly.

Content Extraction

For each query, the SERP was archived; for each search result, the website was archived and the content extracted. To archive the websites, the *Webis Web Archiver* was used (Kiesel et al. [2018]). By archiving the SERP with the first 10 search results for each query, the URLs of the search results and their position on the SERP were saved as well. Using these URLs, the search results

[7] `https://www.startpage.com`
[8] `https://www.startpage.com/en/privacy-policy/?t=default`
[9] `https://www.startpage.com/search/settings?lang=en`

were archived. For each search result, the content of the archived websites was extracted using a content extractor (Kiesel et al. [2017]). Since, the queries of a topic often result in similar search results, URLs, which appeared more than once, were ignored during the archiving and extraction. Based on the nine different topics and three to nine queries per topic, there were 490 search results. Without the duplicates, 350 websites were archived and extracted. Those 350 documents were saved with their corresponding URL and then used for stance classification.

Stance Classification

For each search result document, the stance was inferred using a stance classifier provided by Yamen Ajjour, Bauhaus-Universität Weimar. The stance classifier uses machine learning and trains on data sets from different topics, to classify the stance of a given document on a topic. From the topics used in this experiment, the classifier is trained on the topics *Abortion*, *Death Penalty*, *Gun Control*, *Marijuana Legalization* and *Nuclear Energy*. For those five topics an error rate of 0.3 can be assumed. For *Climate change*, *Health Care*, *Mandatory Vaccination* and *Same Sex Marriage* this is not the case. This has to be kept in mind when looking at the results. The stance classifier works on a sentence level and first identifies the stance of each sentence, this can either be pro, con or neutral. Then, based on the distribution of stances, it assesses the overall stance for the whole document. If 90% of the sentences are neutral, the document is classified as neutral. If there are more pro sentences than con, the document is classified as pro. If there are more con sentences than pro, the document is classified as con. If there is the same amount of pro and con sentences, the document is classified as neutral. Since the classifier has an error rate of 30% and for untrained topics an unknown error rate, the stances need to be spot checked to be able to estimate the impact. In addition to that, there is also the possibility of errors during the archiving and content extraction which leads to problems with the data. That is why, we checked the stances of some documents and classified them ourselves. During the bias assessment, the stances pro, con and neutral are interpreted as 1, -1 and 0.

In table 4.4, an example is provided with the URLs of the topic *Abortion*, the stances classified by the stance classifier and the stances we classified. As can be seen, 6 URLs have been classified wrongly by the classifier, compared to our classification. For the overall 24 URLs this results in an error rate of 25% which is in complete accordance with the expected error rate of the classifier for trained topics. For the topic *Climate Change*, the URLs were classified by us as well. There are 42 URLs overall and 14 were classified wrongly, compared to how we classified them. This leads to an error rate of 33%, and since the classifier is not trained on *Climate Change*, this is still rather close to the expected 30%. The other topics were spot-checked and showed similar tendencies but no overall error-rate was calculated.

Table 4.4: The URLs for the topic *Abortion* with the stances inferred by the classifier and the stances we defined.

URL	Stance Classifier	We
https://www.plannedparenthood.org/learn/abortion	Pro	Pro
https://www.guttmacher.org/gpr/2019/09/us-abortion-rate-continues-drop-once-again-state-abortion-restrictions-are-not-main	Pro	Pro
https://medlineplus.gov/abortion.html	Pro	Neutral
https://www.webmd.com/women/abortion-procedures	Pro	Neutral
https://www.guttmacher.org/state-policy/explore/overview-abortion-laws	Neutral	Neutral
https://www.who.int/health-topics/abortion	Pro	Pro
https://www.wholewomanshealth.com/abortion-care/abortion-options/medication-abortion/	Pro	Pro
https://www.plannedparenthood.org/learn/abortion/the-abortion-pill	Pro	Pro
https://www.health.state.mn.us/people/wrtk/handbook.html	Pro	Neutral
https://www.britannica.com/science/abortion-pregnancy	Pro	Neutral
https://www.plannedparenthood.org/learn/abortion/the-abortion-pill/how-does-the-abortion-pill-work	Pro	Pro
https://nwhn.org/abortion-pills-medication-abortion/	Pro	Pro
https://en.wikipedia.org/wiki/Abortion	Pro	Neutral
https://www.cdc.gov/reproductivehealth/data_stats/abortion.htm	Neutral	Neutral
https://www.amnesty.org/en/what-we-do/sexual-and-reproductive-rights/abortion-facts/	Pro	Pro
https://prochoice.org/patients/abortion-what-to-expect/	Pro	Pro
https://www.theverge.com/2015/8/25/9174769/abortion-pill-shot-surgery-medical-women-healthcare	Pro	Pro
https://www.mayoclinic.org/tests-procedures/medical-abortion/about/pac-20394687	Pro	Pro
https://foundationsoflife.org/facts-about-abortion/	Pro	Con
https://www.healthline.com/health/abortion-pill	Pro	Pro
https://helloclue.com/articles/cycle-a-z/what-to-expect-before-during-and-after-an-abortion	Pro	Pro
https://prochoice.org/	Pro	Pro
https://www.bpas.org/abortion-care/abortion-treatments/the-abortion-pill/abortion-pill-up-to-10-weeks/	Pro	Pro
https://fallschurchhealthcare.com/abortion-pill/	Pro	Pro

Aggregated Stances

Based on the stance $\sigma(d, t)$ of each search result document $d \epsilon D_q$, the aggregated stance $as(q)$ for each query $q \epsilon Q_t$ was calculated (see section 3.2.1). For each topic, the aggregated stances can be seen in table 4.5 and 4.5. In this experiment, ten results are used to calculate the aggregated stance for each query. This leads to a value range of $[-4.54, 4.54]$ for the aggregated stance, based on the formulas of section 3.2.1. A SERP, that only has positive search results, has an aggregated stance of 4.54 and a purely negative SERP, one of -4.54. An aggregated stance of 0 is in this case only possible if all search results are neutral. Therefore, by looking at the sign and the size of the aggregated stance, the overall stance, of the SERP shown for a query, can be inferred.

Example - Abortion:

q_1: abortion
$\sigma(d_i, t)$: 1, 1, 1, 0, 1, 1, 1, 1, 1, 1
$as(q_1) = DCG_{10} = 4.1129$

q_2: abortions
$\sigma(d_i, t)$: 1, 1, 0, 1, 1, 1, 1, 0, 1, 1
$as(q_2) = DCG_{10} = 3.7281$

q_3: pill abortion
$\sigma(d_i, t)$: 1, 1, 1, 1, 1, 1, 1, 1, 1, 1
$as(q_3) = DCG_{10} = 4.5436$

In the example above, it can be seen, that for the three queries, most stances are pro and only a few are neutral. This is correctly reflected in the aggregated stances, since the values are relatively close to and at the upper bound. Further, the influence of neutral results can also be seen, the more neutral results per query, the lower the aggregated stance. The position of the neutral results also influences the results.

In table 4.5 the aggregated stances for the most asked queries of each topic are depicted. Since the aggregated stances differ from topic to topic but are similar within a topic, this shows that the topic dependant classification works. Both *Abortion* and *Gun Control* have relatively high aggregated stance values, at around 3 to 4. This means that the results returned by Google are overall positive towards the topics. For the topic *Marijuana*, most values are also positive and range at around 2. Although, the most asked query portraits the opposite stance at around -2. A reason for this could be that the most asked query is simply *marijuana*, whereas the others contain the word *legal*. Meaning that the other queries are specifically on the legality of marijuana and this seems to result in more positive search results than when asking for marijuana itself. The values of *Mandatory Vaccination* range at around 1 to 2 and seem coherent within the topic. This means that there are more neutral and negative results but overall the stance still remains positive. For the topic *Nuclear Power*, the values are closer to zero, with the exception of two queries containing the word *energy*. This could again be because for *energy* the returned results are about all energy sources and therefore more positive than the results returned for *nuclear power*. The aggregated stance values for *Universal Health Care* are divided by the queries. For the three queries containing the word *uhc* the value is close to and at zero. This is because "uhc" stands for "UnitedHealthcare" and the search results are mainly websites of health care providers or portals to sign into. Therefore such queries are mainly classified as neutral. The other queries, are about universal health care and therefore have a positive stance, ranging from 2 to 3. The values for *Same Sex Marriage* are close to and at zero, which implies overall neutral search results, this is consistent throughout the topic. Both *Climate Change* and *Death Penalty* have very low values, at around -2 to -3, which indicates a lot of search results with negative stances.

Table 4.5: Aggregated stances for the most asked queries of a topic, the queries are ordered from most to least asked query

Topic t	**Most asked queries** $q \epsilon Q_t$	**Aggregated Stances** $as(q)$
Abortion	abortion	4.1129
	abortions	3.7281
	pill abortion	4.5436
Climate Change	climate	-0.0973
	climate change	-3.1552
	warming	-2.4783
	global warming	-3.3311
	climate change is	-3.5421
	global warming is	-2.8634
	what is climate	-0.9592
	global climate	-3.0777
Death Penalty	death penalty	-2.5972
	execution	-1.8362
	the death penalty	-2.6116
	executed	-0.6759
	death row	-0.6309
	executions	-1.2317
	capital punishment	-3.7567
	death penalty states	-3.854
	death sentence	-2.3927
Gun Control	gun	2.8367
	gun laws	3.6057
	gun control	3.9126
Mandatory Vaccinations	is vaccination mandatory	1.6625
	mandatory vaccination bill	1.4307
	mandatory vaccination 2020	1.8332
	mandatory vaccination oklahoma	1.7465
	mandatory vaccination laws	2.0858
	stop mandatory vaccination	1.6023
	mandatory vaccination states	1.3507
	mandatory vaccination bill 2020	0.8755

Table 4.5: Aggregated stances for the most asked queries of a topic, the queries are ordered from most to least asked query (continued)

Topic t	**Most asked queries** $q \epsilon Q_t$	**Aggregated Stances** $as(q)$
Marijuana Legalization	marijuana	-2.3018
	legal marijuana	2.1924
	is marijuana legal	1.7554
	states legal marijuana	2.1666
	legal cannabis	2.3225
Nuclear Power	nuclear	0.0558
	nuclear power	0.5158
	energy	2.6653
	nuclear energy	1.8898
	nuclear plant	0.8379
Same Sex Marriage	gay marriage	0.8333
	gay marriage legal	-0.2891
	gay marriage legalized	-0.1296
	is gay marriage legal	0
	supreme court	0
	gay marriage us	0
	gay wedding	0
Universal Health Care	uhc	1.005
	uhc provider	0.5
	uhc login	0
	universal health	2.1095
	universal healthcare	3.0828
	universal health care	2.8805

4.2.2 Simulated Reference Distribution

The reference distributions $r \epsilon R_t$ of each topic, were then simulated to create the simulated reference distribution P_r and the sample S_r, to be used during the bias assessment. This was done, as described in section 3.2.2, simulating the stances of ten search results and repeating it a 1000 times for each $r \epsilon R_t$.

Example - Abortion:

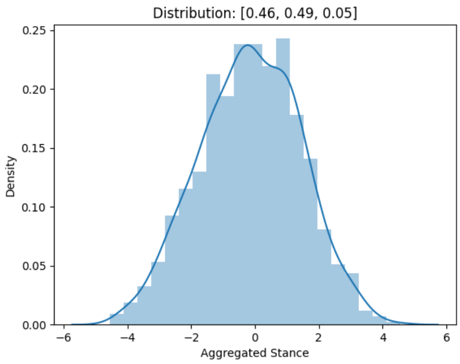

Figure 4.1: Graph of the simulated reference distribution $P_{[0.46,0.49,0.05]}$ (Opinion Poll) for *Abortion*

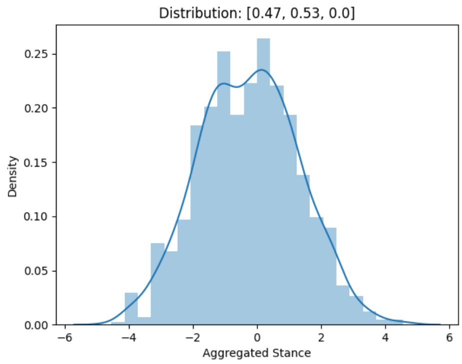

Figure 4.2: Graph of the simulated reference distribution $P_{[0.47,0.53,0.0]}$ (Political Landscape) for *Abortion*

In Figure 4.1 and 4.2 the graphs of the simulated reference distributions for the topic *Abortion* are shown. On the x-axis the range of aggregated stances is depicted and on the y-axis the kernel density, meaning the estimated probabilities of the aggregated stances, can be seen. In Figure 4.1 the graph, for the simulated reference distribution $P_{[0.46,0.49,0.05]}$ out of the category *Opinion Poll*, is depicted. The aggregated stances in the graph range from about -4.8 to 4.5 and since the reference distribution consists of 46% pro and 49% con, the graph correctly depicts this difference. In Figure 4.2 the graph, for the simulated reference distribution $P_{[0.47,0.53,0.0]}$ out of the category *Political Landscape* can be seen. The distribution is similar to the one from Figure 4.1, with 47% pro and 53% con, this is again correctly depicted by the the graph.

The graphs of the other simulated reference distributions, for each topic, are shown in the appendix.

4.2.3 Bias Assessment

For the bias assessment, first the probabilities for the individual queries and then for the whole topic are calculated.

Query Probability

As a first step of the bias assessment, the probability $p_r(as(q))$ of the aggregated stance $as(q)$ of each query $q \epsilon Q_t$, belonging to each simulated reference distribution P_r, for each $r \epsilon R_t$, was calculated. For this a permutation test, as explained in section 3.2.3 was used. In the following an example of the probabilities calculated for the topic *Abortion* is provided.

Example - Abortion:

$q_1 : p_{[0.46,0.49,0.05]}(4.1129) = 0.002$
$q_1 : p_{[0.47,0.53,0.0]}(4.1129) = 0.002$

$q_2 : p_{[0.46,0.49,0.05]}(3.7281) = 0.007$
$q_2 : p_{[0.47,0.53,0.0]}(3.7281) = 0.006$

$q_3 : p_{[0.46,0.49,0.05]}(4.5436) = 0.0$
$q_3 : p_{[0.47,0.53,0.0]}(4.5436) = 0.0$

For the topic *Abortion* there are three most asked queries (q_1, q_2, q_3) and two reference distributions $([0.46, 0.49, 0.05], [0.47, 0.53, 0.0])$. Because the probability $p_r(as(q))$ is calculated for each query and for each reference distribution, there are 6 different probabilities calculated for the topic *Abortion*. As can be seen, the probabilities are all close to zero. This means that the queries most likely do not belong to the expected reference distributions. This can also be seen in Figure 4.3 and 4.4. Here the aggregated stances of the three queries are marked in the graphs of the two reference distributions. As can be seen, the aggregated stance values of the queries are at the edge and outside of the distribution, this corresponds with the calculated probabilities and shows the miss-match between the queries and the reference distributions. This already implies a bias of Google for *Abortion* as Google returns a far more positive view on *Abortion* than reflected in the opinion of the population or in the Senate.

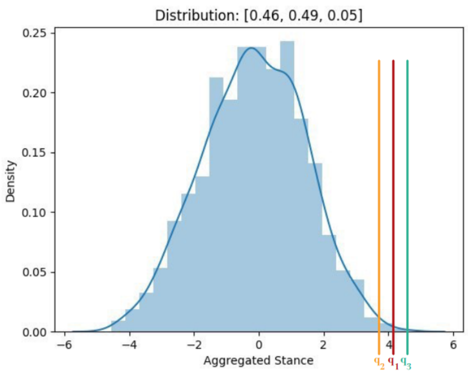

Figure 4.3: Graph of the simulated reference distribution $P_{[0.46,0.49,0.05]}$ for *Abortion* with marks for the aggregated stances of the queries q_1, q_2 and q_3

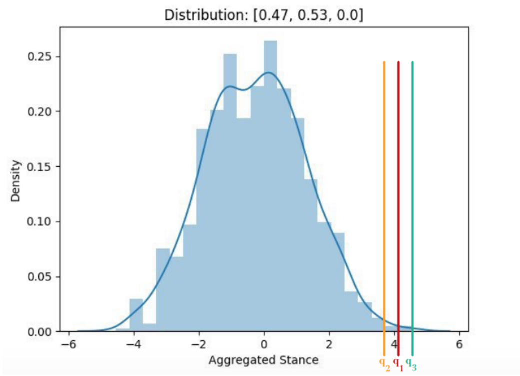

Figure 4.4: Graph of the simulated reference distribution $P_{[0.47,0.53,0.0]}$ with marks for the aggregated stances of the queries q_1, q_2 and q_3

Topic Probability

To calculate the probability $p_r(t)$ of the whole topic belonging to the simulated reference distribution, the probabilities $p_r(as(q))$ of all queries $q \epsilon Q_t$ were weighted according to the query weight $w(q)$ and then summed up. Since we assume to cover about 90% of all user inquiries with our top queries, we account for the leftover 10% by giving *Google* the benefit of the doubt and assuming the expected value of the distribution as the aggregated stance. Therefore the summed up probabilities are weighted with 0.9 and the probability of the expected value, 1.0, is multiplied by 0.1 and added to the calculated value. This results in $p_r(t)$. $p_r(t)$ is the probability of an arbitrary Google user in the US, during 2019-2020, seeing search results, on a given topic, with a ranking that reflects the expected reference distribution. Because of the benefit of the doubt, the resulting probabilities of the topics are always at least 0.1.

Example - Abortion:

$$p_{opinion}(Abortion) = 0.9 * 0.00249 + 0.1 * 1.0 = 0.1022$$
$$p_{political}(Abortion) = 0.9 * 0.00236 + 0.1 * 1.0 = 0.1021$$

The calculated probabilities for *Abortion* are very similar for both simulated reference distributions since they started out with a similar pro, con and neutral distribution. Each probability portraits how likely it is that the ranking of the search results, as returned by Google for the topic *Abortion*, reflect the reference distribution. Meaning the probabilities show, how likely it is that

Table 4.6: Probability $p_r(t)$ for each topic and reference distribution, sorted by the categories *Opinion Poll*, *Market Information* and *Political Landscape*

Topic t	**Opinion Poll** $p_{r_1}(t)$	**Market Information** $p_{r_2}(t)$	**Political Landscape** $p_{r_3}(t)$
Abortion	0.1022		0.1021
Climate Change	0.5702		0.3447
Death Penalty	0.1620	0.1927	
Gun Control	0.4020	0.3583	0.1198
Mandatory Vaccinations	0.5887	0.1	
Marijuana Legalization	0.5410	0.2960	
Nuclear Power	0.8241	0.1	
Same Sex Marriage	0.6137		0.7158
Universal Health Care	0.6197		0.4665

there is no bias. To conclude, for the topic *Abortion* the result is as follows: The probability of an arbitrary Google user, in the US during 2019-2020, seeing search results on the topic *Abortion* with a ranking that reflects the expected reference distributions, is about 0.1. Since this probability is close to 0.1, it is very unlikely that there is no bias.

In table 4.6 the probabilities of all topics belonging to the corresponding expected distributions can be seen. The probabilities for a topic can differ for each reference distribution, therefore the assessed bias is ambiguous and can be interpreted multiple ways. In the following, the probabilities of the topics are separated by reference distribution category and displayed in individual tables (see tables 4.7, 4.8 and 4.9). This way a separate bias assessment is possible.

Table 4.7: Probability $p_r(t)$ for each topic having a reference distribution, out of the category *Opinion Poll*; probabilities sorted from least to most

Topic t	**Opinion Poll** $p_{r_1}(t)$
Abortion	0.1022
Death Penalty	0.1620
Gun Control	0.4020
Marijuana Legalization	0.5410
Climate Change	0.5702
Mandatory Vaccinations	0.5887
Same Sex Marriage	0.6137
Universal Health Care	0.6197
Nuclear Power	0.8241

In table 4.7, the probabilities for all nine topics and the corresponding reference distributions out of the category *Opinion Poll* are shown. Here, the calculated probabilities vary from about 0.1 up to 0.8 but overall the probabilities are rather large, in six out of nine cases they are above 0.5. This indicates that for most topics a bias is unlikely, especially for *Nuclear Power* as this topic has a probability of 0.8. For the topics *Abortion* and *Death Penalty*, on the other hand, it is unlikely that there is no bias, as they both have low probabilities at around 0.1 to 0.2.

Table 4.8: Probability $p_r(t)$ for each topic having a reference distribution, out of the category *Market Information*; probabilities sorted from least to most

Topic t	**Market Information** $p_{r_2}(t)$
Mandatory Vaccinations	0.1
Nuclear Power	0.1
Death Penalty	0.1927
Marijuana Legalization	0.2960
Gun Control	0.3583

In table 4.8, the probabilities for the five topics with a corresponding reference distribution out of the category *Market Information* are shown. Here, the calculated probabilities are all relatively small, ranging from 0.1 to about 0.35. This indicates that for those five topics, the possibility that there is no bias is very low, especially for *Mandatory Vaccination* and *Nuclear Power*, since the probabilities are exactly 0.1. This is the lowest possible probability and indicates that it is unlikely that there is no bias.

Table 4.9: Probability $p_r(t)$ for each topic having a reference distribution, out of the category *Political Landscape*; probabilities sorted from least to most

Topic t	**Political Landscape** $p_{r_3}(t)$
Abortion	0.1021
Gun Control	0.1198
Climate Change	0.3447
Universal Health Care	0.4665
Same Sex Marriage	0.7158

In table 4.9, the probabilities for the five topics with a corresponding reference distribution out of the category *Political Landscape* are shown. The probabilities range from about 0.1 to 0.7, with four out of five being below 0.5. Here, the topics *Abortion* and *Gun Control* have a probability of about 0.1, implicating that it is unlikely that there is no bias. *Same Sex Marriage* on the other hand, has a probability of 0.7, making it likely that there is no bias.

When comparing the probabilities for the different reference distributions of one topic, further observations can be made (see table 4.6). For *Abortion* and *Death Penalty*, the reference distributions out of the different categories all result in similar probabilities. For *Abortion*, the probabilities lie at about 0.1 and for *Death Penalty* they are all below 0.2. Since the reference distributions match, a bias can be assumed for those two cases. For *Mandatory Vaccination* and *Nuclear Power*, there are large differences between the probabilities calculated for the reference distributions out of the categories *Opinion Poll* and *Market Information*. For the category *Market Information*, the probabilities are 0.1, for both topics, making it unlikely that there is no bias. For the category *Opinion Poll* on the other hand, the probabilities are far larger, in comparison, at about 0.6 for *Mandatory Vaccination* and 0.8 for *Nuclear Power*. This would mean that it is likely that there is no bias. As described in section 4.1.3, the reference distributions out of the category *Market Information* could not be correctly mapped to the pro, con and neutral scheme used for the bias assessment. This transfer of statistical data into an opinion related scheme is not entirely correct and could be the reason for these differences. For *Same Sex Marriage*, the calculated probabilities are relatively high for both reference distributions, at around 0.6 and 0.7. This indicates that there is no bias for this topic.

Chapter 5

Conclusion and Future Work

Search engines are a part of everyday life, yet only few users realize the influence they can have. Search engines are used for accessing information, finding solutions to problems and reaching decisions. Since the internet offers, what feels like, unlimited information about every possible topic, search engines act as a type of filter. It is widely known that in response to a query, they offer those results that most closely match it and are of most relevance. Something that the user is not always aware of, is the role and possible influence of the search engine in this process. To counteract this, an approach for assessing bias in search engines was developed in this thesis. In this approach, the bias of one search engine for specific topics and a selected location is studied. In each step of the approach, the focus lies on detecting bias that has the potential to influence society. This approach was successfully used in an experiment, to examine the bias of Google for nine topics, during 2019-2020 in the US. Based on the results of the stance classifier and the reference distributions, it could be shown in the experiment that a bias for the two topics *Abortion* and *Death Penalty* is likely. For the topic *Same Sex Marriage* it could be shown that a bias is unlikely.

In the future, the approach could be adapted to include multiple search engines. Using the same topics and location, the resulting biases could be compared. This way, the behaviour of different search engines could be studied and more insight into the bias for one location could be gained. Another possibility would be adapting the approach to include multiple locations, for the bias assessment of one search engine and specific topics. This way, the behaviour of this search engine and its bias could be studied further. In the approach, a stance classifier was used to classify the stances of the search results. Since stance classifiers are error prone, there is the possibility of forgoing this altogether and using crowd sourcing instead. This might lead to a more reliable stance classification but it would also take more time and could not be operationalized.

The error rate of the stance classifier used in the experiment was 0.3. This distorts and influences the bias assessment. In the future and with changes in technology, stance classifiers with lower error rates can be used. This would improve the accuracy of the bias assessment. Further, the used classifier was only trained on data sets from five out of the nine topics of the experiment. For the four untrained topics, the error rate of 0.3 can not be assumed. This leads to further distortions in the bias assessment. In future works, a stance classifier trained on all topics should be used, to avoid such errors. During the experiment, the bias was assessed for nine topics. In the future, more topics could be added, to gain more knowledge about the bias of Google. The reference distributions are adaptable as well and should be expanded to create more resources for the bias assessment. During the bias assessment the search engine Startpage was used instead of Google. This has multiple advantages, like a constant amount of search results and no adaptation to user behaviour. However, when assessing the bias of Google in the future, directly using the SERP of Google might lead to a more accurate bias assessment. Since Google highlights some results and adds their own services on top of the SERP, it would be more accurate to consider this as well. Although then, a method for interpreting the Google result page would need to be developed and the weighting of the search results would need to be adapted. This is, because user favour such highlighted search results over others and perceive them differently.

This thesis described how an accurate bias assessment of search engines can look like and successfully showed its utilisation. Multiple options for expanding the described approach in the future, to enhance the bias assessment, have been shown.

Appendix A

Simulated Reference Distributions

Abortion

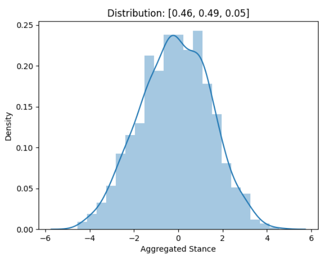

Figure A.1: $P_{[0.46,0.49,0.05]}$ *(Opinion Poll)*

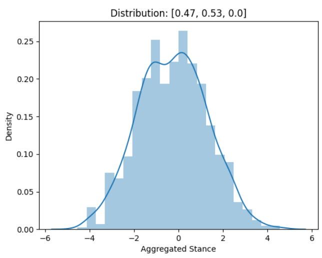

Figure A.2: $P_{[0.47,0.53,0.0]}$ *(Political Landscape)*

Climate Change

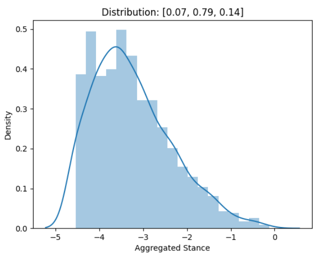

Figure A.3: $P_{[0.07,0.79,0.14]}$ *(Opinion Poll)*

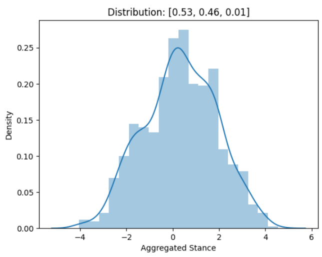

Figure A.4: $P_{[0.53,0.46,0.01]}$ *(Political Landscape)*

Death Penalty

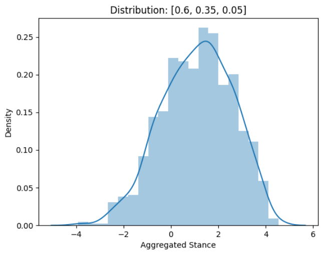

Figure A.5: $P_{[0.6,0.35,0.05]}$ *(Opinion Poll)*

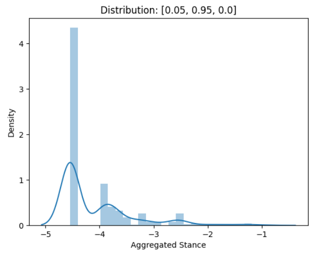

Figure A.6: $P_{[0.05,0.95,0.0]}$ *(Market Information)*

Gun Control

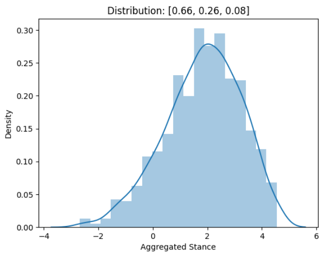

Figure A.7: $P_{[0.66,0.26,0.08]}$ *(Opinion Poll)*

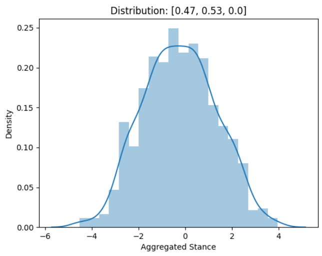

Figure A.8: $P_{[0.47,0.53,0.0]}$ *(Political Landscape)*

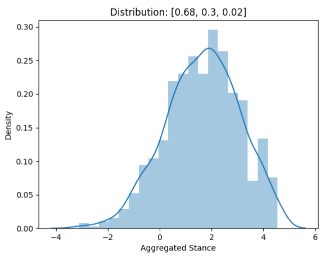

Figure A.9: $P_{[0.68,0.3,0.02]}$ *(Market Information)*

Mandatory Vaccination

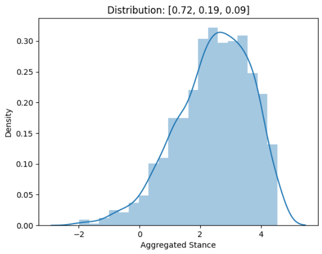

Figure A.10: $P_{[0.72,0.19,0.09]}$ *(Opinion Poll)*

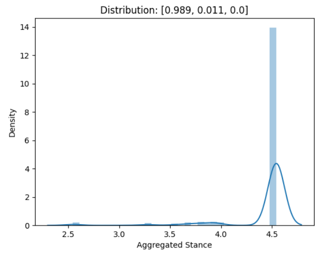

Figure A.11: $P_{[0.99,0.01,0.0]}$ *(Market Information)*

Marijuana Legalization

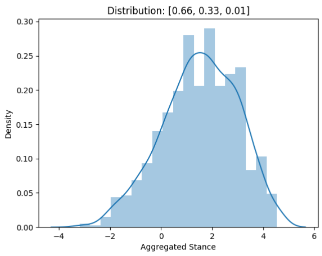

Figure A.12: $P_{[0.66,0.33,0.01]}$ *(Opinion Poll)*

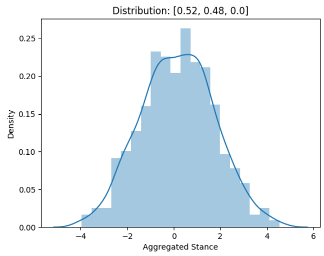

Figure A.13: $P_{[0.52,0.48,0.0]}$ *(Market Information)*

Nuclear Power

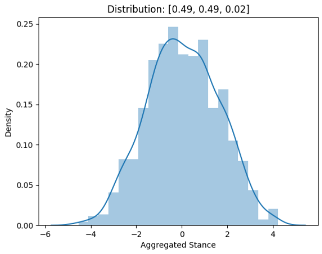

Figure A.14: $P_{[0.49,0.49,0.02]}$ *(Opinion Poll)*

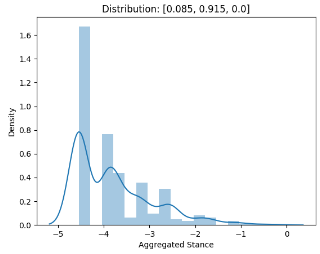

Figure A.15: $P_{[0.09,0.91,0.0]}$ *(Market Information)*

Same Sex Marriage

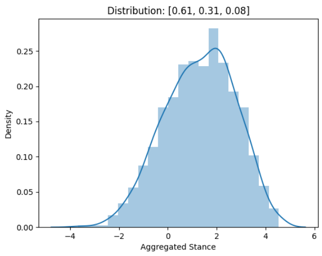

Figure A.16: $P_{[0.61,0.31,0.08]}$ *(Opinion Poll)*

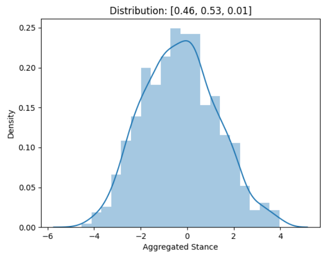

Figure A.17: $P_{[0.46,0.53,0.01]}$ *(Political Landscape)*

Universal Health Care

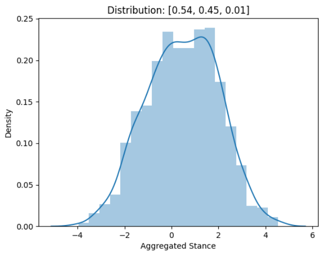

Figure A.18: $P_{[0.54,0.45,0.01]}$ *(Opinion Poll)*

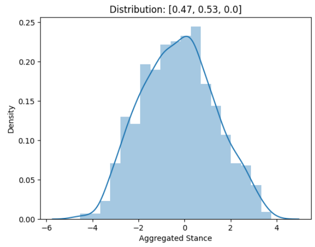

Figure A.19: $P_{[0.47,0.53,0.0]}$ *(Political Landscape)*

Bibliography

U.S. Energy Information Administration. Total energy. `https://www.eia.gov/totalenergy/data/browser/index.php?tbl=T01.03#/?f=A`, 2019. accessed November 1, 2020. 4.2

R. Baeza-Yates, C. Hurtado, M. Mendoza, and G. Dupret. Modeling user search behavior. In *Third Latin American Web Congress (LA-WEB'2005)*, 2005. DOI: `https://doi.org/10.1109/LAWEB.2005.23`. 4.2.1

Ing-Xiang Chen and Cheng-Zen Yang. Position Paper: A Study of Web Search Engine Bias and its Assessment. In *IW3C2 WWW*, 2006. 2.2, 2.1

Morning Consult and Politico. National tracking poll #190826. `https://morningconsult.com/wp-content/uploads/2019/08/190826_crosstabs_POLITICO_RVs_v2_JB.pdf`, August 2019. page 148, accessed November 1, 2020. 4.2

Elizabeth Davis and Tracy L. Snell. Capital punishment, 2016. `https://www.bjs.gov/content/pub/pdf/cp16sb.pdf`, April 2018. accessed November 1, 2020. 4.2

Diffen. Democrat vs. republican. `https://www.diffen.com/difference/Democrat_vs_Republican`, October 2020. accessed November 1, 2020. 4.2

Benjamin Edelman. Bias In Search Results: Diagnosis And Response. In *The Indian Journal Of Law And Technology Vol.7*, 2011. 2.1

Robert Epstein and Ronald E. Robertson. The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. In *PNAS*, 2015. DOI: `https://doi.org/10.1073/pnas.1419828112`. 1

Gallup. Moral issues - gallup historical trends. `https://news.gallup.com/poll/1681/moral-issues.aspx`, May 2019a. accessed November 1, 2020. 4.2

Gallup. Healthcare system. https://news.gallup.com/poll/4708/healthcare-system.aspx, 2019b. accessed November 1, 2020. 4.2

Gallup. Illegal drugs - gallup historical trends. https://news.gallup.com/poll/1657/Illegal-Drugs.aspx, October 2019c. accessed November 1, 2020. 4.2

Eric Goldman. Search Engine Bias And The Demise Of Search Engine Utopianism. In *8 Yale J.L. & Tech*, 2006. 2.1

HA Hill, LD Elam-Evans, D Yankey, JA Singleton, and Y Kang. Vaccination Coverage Among Children Aged 19-35 Months - United States. In *MMWR Morb Mortal Wkly Rep 2018;67:1123–1128*, 2017. DOI: http://dx.doi.org/10.15585/mmwr.mm6740a4. 4.2

Mark T. Keane, Maeve O'Brien, and Barry Smyth. Are people biased in their use of search engines? In *Communications of the ACM*, 2008. DOI: https://doi.org/10.1145/1314215.1314224. 1

Johannes Kiesel, Benno Stein, and Stefan Lucks. A Large-scale Analysis of the Mnemonic Password Advice. In *24th Annual Network and Distributed System Security Symposium (NDSS 2017)*, 2017. DOI: https://doi.org/10.14722/ndss.2017.23077. 4.2.1

Johannes Kiesel, Florian Kneist, Milad Alshomary, Benno Stein, Matthias Hagen, and Martin Potthast. Reproducible Web Corpora: Interactive Archiving with Automatic Quality Assessment. In *ACM J. Data Inform. Quality 0, 0, Article 0 (August 2018)*, 2018. DOI: https://doi.org/10.1145/3239574. 4.2.1

Juhi Kulshrestha, Motahhare Eslami, Johnnatan Messias, Muhammad Bilal Zafar, Saptarshi Ghosh, Krishna Gummadi, and Karrie Karahalios. Quantifying Search Bias: Investigating Sources of Bias for Political Searches in Social Media. In *CSCW '17: Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 417–432, 2017. DOI: https://doi.org/10.1145/2998181.2998321. 2.2, 2.1

Juhi Kulshrestha, Motahhare Eslami, Johnnatan Messias, Muhammad Bilal Zafar, Saptarshi Ghosh, Krishna Gummadi, and Karrie Karahalios. Search bias quantification: investigating political bias in social media and web search. In *Information Retrieval Journal 22*, pages 188–227, 2019. DOI: https://doi.org/10.1007/s10791-018-9341-2. 2.2, 2.1

Dirk Lewandowski. Is Google Responsible For Providing Fair And Unbiased Results? In *Taddeo M., Floridi L. (eds) The Responsibilities of Online Service Providers. Law, Governance and Technology Series, vol 31. Springer, Cham*, 2017. DOI: `https://doi.org/10.1007/978-3-319-47852-4_4`. 2.1

MaristPoll. Yahoo news/marist poll: Weed & the american family. `http://maristpoll.marist.edu/wp-content/misc/Yahoo%20News/20170417_Methodology%20and%20Nature%20of%20the%20Sample_Yahoo%20News-Marist%20Poll.pdf`, March 2017. accessed November 1, 2020. 4.2

Abbe Mowshowitz and Akira Kawaguchi. Assessing bias in search engines. In *Information Processing and Management 38*, pages 141–156, 2002. DOI: `https://doi.org/10.1016/S0306-4573(01)00020-6`. 2.2, 2.1

Yusra Murad. As states rethink vaccine laws, support for mandatory shots lowest with young adults. `https://morningconsult.com/2019/05/15/as-states-rethink-vaccine-laws-support-for-mandatory-shots-lowest-with-young-adults/`, May 2019. accessed November 1, 2020. 4.2

Alamir Novin and Eric Meyers. Making Sense of Conflicting Science Information: Exploring Bias in the Search Engine Result Page. In *CHIIR '17: Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, pages 175–184, 2017. DOI: `https://doi.org/10.1145/3020165.3020185`. 1

Bing Pan, Helene Hembrooke, Thorsten Joachims, Lori Lorigo, Geri Gay, and Laura Granka. In Google We Trust: Users' Decisions on Rank, Position, and Relevance. In *Journal of Computer-Mediated Communication, Volume 12, Issue 3*, pages 801–823, 2007. DOI: `https://doi.org/10.1111/j.1083-6101.2007.00351.x`. 1

Kim Parker, Juliana Menasce Horowitz, Ruth Igielnik, J. Baxter Oliphant, and Anna Brown. America's complex relationship with guns. `https://www.pewsocialtrends.org/2017/06/22/americas-complex-relationship-with-guns/`, June 2017. accessed November 1, 2020. 4.2

PewResearchCenter. Attitudes on same-sex marriage. `https://www.pewforum.org/fact-sheet/changing-attitudes-on-gay-marriage/`, May 2019. accessed November 1, 2020. 4.2

Evaggelia Pitoura, Panayiotis Tsaparas, Giorgos Flouris, Irini Fundulaki, Panagiotis Papadakos, Serge Abiteboul, and Gerhard Weikum. On Measuring Bias in Online Information. In *ACM SIGMOD Record, Vol. 46*, 2018. DOI: `https://doi.org/10.1145/3186549.3186553`. 2.1, 2.2, 2.1

Rj Reinhart. 40 years after three mile island, americans split on nuclear power. `https://news.gallup.com/poll/248048/years-three-mile-island-americans-split-nuclear-power.aspx`, March 2019. accessed November 1, 2020. 4.2

Lydia Saad. Majority in u.s. still want abortion legal, with limits. `https://news.gallup.com/poll/259061/majority-abortion-legal-limits.aspx`, June 2019. accessed October 25, 2020. 4.2

Jonathan D. Salant. 10 huge differences between democratic and republican platforms. `https://www.nj.com/politics/2016/07/dnc_2016_10_big_ways_the_democratic_platform_diffe.html`, July 2019. accessed November 1, 2020. 4.2

Votesmart. Angus king, jr.'s issue positions (political courage test). `https://justfacts.votesmart.org/candidate/political-courage-test/22381/angus-king-jr`, 2018. accessed November 1, 2020. 4.2

Votesmart. Bernie sanders' issue positions (political courage test). `https://justfacts.votesmart.org/candidate/political-courage-test/27110/bernie-sanders`, 2020. accessed November 1, 2020. 4.2

Karsten Weber. Search Engine Bias. In *AKA Verlag Heidelberg*, 2011. 2.1

Ryen W. White. Beliefs and Biases in Web Search. In *SIGIR '13: Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 3–12, 2013. DOI: `https://doi.org/10.1145/2484028.2484053`. 1

YouGov and TheEconomist. The economist/yougov poll, march 10 - 12, 2019 - 1500 us adult citizens. `https://d25d2506sfb94s.cloudfront.net/cumulus_uploads/document/a5islr4zij/econTabReport.pdf`, March 2019. page 30, accessed November 1, 2020. 4.2