Quantifying Search Bias: Investigating Sources of Bias for Political Searches in Social Media

Juhi Kulshrestha

joint work with

Motahhare Eslami, Johnnatan Messias, Muhammad Bilal Zafar, Saptarshi Ghosh, Krishna P. Gummadi and Karrie Karahalios

MAX PLANCK INSTITUTE FOR SOFTWARE SYSTEMS
Social media as "search" platform

- Rich source of news & information
- **Search** is how users follow news about events & people
- **Hashtags** - recommended queries
Social media as "search" platform

#obamacare
Social media as "search" platform

Ranked list (according to importance)
Potential bias in search results

#obamacare

Mark Meadows
@RepMarkMeadows

I support @RandPaul and @RepSanfordSC's #Obamacare replacement plan -- a plan that will lower costs and put the focus back on the patient.
Potential bias in search results

Mark Meadows
@RepMarkMeadows

I support @RandPaul and @RepSanfordSC's #Obamacare replacement plan -- a plan that will lower costs and put the focus back on the patient.

Ryan
@Politicalry

Support for Obamacare growing! We do not need to #RepealAndReplace #Obamacare, fix it as is or keep it for the sake of the people's welfare.
Search can shape user opinion

- Users place \textit{greater trust in higher ranked items} [Pan et al., 2007]
- Biased search results can \textit{influence voting patterns} [Epstein & Robertson, 2015]
Search bias in the headlines

Search engine bias: What search results are telling you (and what they're not)

How Google Shapes the News You See About the Candidates

Who would Google vote for? An analysis of political bias in internet search engine results

How Google's search algorithm spreads false information with a rightwing bias

Donald Trump Accuses Google of Bias in Search Engine Results
Search bias in the headlines

Search engine bias: What search results are telling you (and what they're not)

How Google Shapes the News You See About the Candidates

Who would Google vote for? An analysis of political bias in internet search engine results

Donald Trump Accuses Google of Bias in Search Engine Results

How Google's search algorithm spreads false information with a rightwing bias

What does bias of a search system mean? How can we quantify it?
Identify sources of search bias

Quantify bias of each source

Study bias of political searches in Twitter
Identify sources of search bias

Quantify bias of each source

Study bias of political searches in Twitter
Social media search engine design

Data corpus
Social media search engine design

Data corpus

q

i_1
i_2
i_3
...
(set)

input
(relevant items)
Social media search engine design

Data corpus → q → i₁, i₂, i₃, ... (set) → Ranking system

input (relevant items) → ranking
Social media search engine design

Data corpus \rightarrow q \rightarrow i_1, i_2, i_3, \ldots \text{(set)} \rightarrow \text{Ranking system} \rightarrow i_5, i_1, i_3, \ldots \text{(output)}

input (relevant items) \rightarrow \text{ranking} \rightarrow \text{output (ranked list)}
Social media search engine design

Data corpus \( \xrightarrow{q} \) i_1, i_2, i_3, ... (set) \( \xrightarrow{\text{Ranking system}} \) i_5, i_1, i_3, ... (set)

input (relevant items) ranking output (ranked list)
Social media search engine design

Output bias may stem from
- Bias introduced by the ranking system
Output bias may stem from
- Bias introduced by the ranking system
- Bias in the input relevant item set
Identify sources of search bias

Quantify bias of each source

Study bias of political searches in Twitter

@juhi153

Quantifying Search Bias

CSCW 2017
Quantifying bias of each source

Step 1: Quantify bias of an individual item
Step 1: Quantifying bias of a single items
Step 1: Quantifying bias of a single item

- We use **source bias** as a proxy
  - Infer bias of each individual item from the bias of the author
  - High scalability
We use source bias as a proxy to infer bias of each individual item from the bias of the author. This approach offers high scalability. Prior work on inferring content bias could be plugged into our bias quantification framework.
Step 1: Quantifying bias of a single item

Evaluation: High agreement between source and content bias (75% or more)
Step 1: Quantifying bias of a single item

Evaluation: High agreement between source and content bias (75% or more)

Mark Meadows
@RepMarkMeadows

I support @RandPaul and @RepSanfordSC's #Obamacare replacement plan -- a plan that will lower costs and put the focus back on the patient.
Step 1: Quantifying bias of a single item

Evaluation: High agreement between source and content bias (75% or more)

Mark Meadows
@RepMarkMeadows
"R-NC 11th District, Republican party"

I support @RandPaul and @RepSanfordSC's #Obamacare replacement plan -- a plan that will lower costs and put the focus back on the patient.
Step 1: Quantifying bias of a single item

Evaluation: High agreement between source and content bias (75% or more)

Mitt Romney
@MittRomney

If Trump had said 4 years ago the things he says today about the KKK, Muslims, Mexicans, disabled, I would NOT have accepted his endorsement.
Step 1: Quantifying bias of a single item

Evaluation: High agreement between source and content bias (75% or more)

If Trump had said 4 years ago the things he says today about the KKK, Muslims, Mexicans, disabled, I would NOT have accepted his endorsement.
Quantifying bias of each source

Data corpus \( \Rightarrow q \Rightarrow \{i_1, i_2, i_3, \ldots\} \Rightarrow \text{Ranking system} \Rightarrow \{i_5, i_1, i_3, \ldots\}

input (relevant items) \rightarrow \text{ranking} \rightarrow \text{output (ranked list)}

Step 2: Quantify bias of a set of items
Step 2: Quantifying bias of set of items

- Compute bias score \( s \) for each item
- Take the average over the whole set

\[
IB(q) = \frac{\sum_{i=1}^{n} s_i}{n}
\]
Quantifying bias of each source

Data corpus → $q$ → $i_1 \ i_2 \ i_3 \ ...$ (set) → Ranking system → $i_5 \ i_1 \ i_3 \ ...$ (set)

Step 3: Quantify bias of a ranked list of items
Step 3: Quantifying bias of ranked list of items

- MAP-style measure
- Bias till rank \( r \)
- Output bias

\[
B(q, r) = \frac{\sum_{i=1}^{r} s_i}{r}
\]

\[
OB(q, r) = \frac{\sum_{i=1}^{r} B(q, i)}{r}
\]
Quantifying bias of each source

Step 4: Quantify bias introduced by ranking system
Step 4: Quantifying bias introduced by ranking

Ranking bias = Output bias - Input bias
Identify sources of search bias

Quantify bias of each source

Study bias of political searches in Twitter
Studying bias for political searches in Twitter

• Search queries for
  • 2016 Democratic and Republican presidential primary debates (#demdebate, rep debate, ...)
  • Presidential candidates (Hillary Clinton, Donald Trump, ...)


Studying bias for political searches in Twitter

- **Search queries for**
  - 2016 Democratic and Republican presidential primary debates (*#demdebate, rep debate, ...*)
  - Presidential candidates (*Hillary Clinton, Donald Trump, ...*)

- **Data collected**
  - **Output**: Twitter top search snapshots every 10 mins
  - **Input**: All tweets containing the query, using streaming api
Studying bias for political searches in Twitter

• Search queries for
  • 2016 Democratic and Republican presidential primary debates (#demdebate, rep debate, ...)
  • Presidential candidates (Hillary Clinton, Donald Trump, ...)

• Data collected
  • **Output**: Twitter top search snapshots every 10 mins
  • **Input**: All tweets containing the query, using streaming api

Non-personalized search data
Studying bias for political searches in Twitter

- Search queries for
  - 2016 Democratic and Republican presidential primary debates (#demdebate, rep debate, ...)
  - Presidential candidates (Hillary Clinton, Donald Trump, ...)

- Data collected
  - Output: Twitter top search snapshots every 10 mins
  - Input: All tweets containing the query, using streaming api

- Computed input, ranking, and output bias for each of the 25 queries
Bias in Twitter search: Input bias vs. Ranking bias

- Does input bias matter?
## Impact of input bias

<table>
<thead>
<tr>
<th>Query</th>
<th>Output Bias</th>
<th>Input Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bernie Sanders</td>
<td>0.71</td>
<td>0.55</td>
</tr>
<tr>
<td>Martin O'Malley</td>
<td>0.64</td>
<td>0.57</td>
</tr>
<tr>
<td>Rand Paul</td>
<td>-0.37</td>
<td>-0.18</td>
</tr>
<tr>
<td>John Kasich</td>
<td>-0.09</td>
<td>-0.13</td>
</tr>
<tr>
<td>dem debate</td>
<td>0.52</td>
<td>0.29</td>
</tr>
<tr>
<td>#demdebate</td>
<td>0.57</td>
<td>0.56</td>
</tr>
<tr>
<td>republican debate</td>
<td>0.53</td>
<td>0.27</td>
</tr>
<tr>
<td>rep debate</td>
<td>0.31</td>
<td>0.40</td>
</tr>
</tbody>
</table>
### Impact of input bias

<table>
<thead>
<tr>
<th>Query</th>
<th>Output Bias</th>
<th>Input Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bernie Sanders</td>
<td>0.71</td>
<td>0.55</td>
</tr>
<tr>
<td>Martin O'Malley</td>
<td>0.64</td>
<td>0.57</td>
</tr>
<tr>
<td>Rand Paul</td>
<td>-0.37</td>
<td>-0.18</td>
</tr>
<tr>
<td>John Kasich</td>
<td>-0.09</td>
<td>-0.13</td>
</tr>
<tr>
<td>dem debate</td>
<td>0.52</td>
<td>0.29</td>
</tr>
<tr>
<td>#demdebate</td>
<td>0.57</td>
<td>0.56</td>
</tr>
<tr>
<td>republican debate</td>
<td>0.53</td>
<td>0.27</td>
</tr>
<tr>
<td>rep debate</td>
<td>0.31</td>
<td>0.40</td>
</tr>
</tbody>
</table>

**Input bias matters!**
### Impact of input bias

<table>
<thead>
<tr>
<th>Query</th>
<th>Output Bias</th>
<th>Input Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bernie Sanders</td>
<td>0.71</td>
<td>0.55</td>
</tr>
<tr>
<td>Martin O'Malley</td>
<td>0.64</td>
<td>0.57</td>
</tr>
<tr>
<td>Rand Paul</td>
<td>-0.37</td>
<td>-0.18</td>
</tr>
<tr>
<td>John Kasich</td>
<td>-0.09</td>
<td>-0.13</td>
</tr>
<tr>
<td>dem debate</td>
<td>0.52</td>
<td>0.29</td>
</tr>
<tr>
<td>#demdebate</td>
<td>0.57</td>
<td>0.56</td>
</tr>
<tr>
<td>republican debate</td>
<td>0.53</td>
<td>0.27</td>
</tr>
<tr>
<td>rep debate</td>
<td>0.31</td>
<td>0.40</td>
</tr>
</tbody>
</table>

Input bias varies across queries
<table>
<thead>
<tr>
<th>Query</th>
<th>Input Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>dem debate</td>
<td>0.29</td>
</tr>
<tr>
<td>#demdebate</td>
<td>0.56</td>
</tr>
<tr>
<td>republican debate</td>
<td>0.27</td>
</tr>
<tr>
<td>rep debate</td>
<td>0.40</td>
</tr>
</tbody>
</table>
Effect of query phrasing

<table>
<thead>
<tr>
<th>Query</th>
<th>Input Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>dem debate</td>
<td>0.29</td>
</tr>
<tr>
<td>#demdebate</td>
<td>0.56</td>
</tr>
<tr>
<td>republican debate</td>
<td>0.27</td>
</tr>
<tr>
<td>rep debate</td>
<td>0.40</td>
</tr>
</tbody>
</table>

Even for the same event, query phrasing can greatly effect the bias.
Bias in Twitter search: Input bias vs. Ranking bias

- Does input bias matter?
  - Input bias does matter
  - Can vary significantly based on the query
  - Even for the same event, different phrasings of queries have widely differing biases
Bias in Twitter search: Input bias vs. Ranking bias

• Does input bias matter?
  • Input bias does matter
  • Can vary significantly based on the query
    • Even for the same event, different phrasings of queries have widely differing biases

• Does the ranking bias exist?
## Examining ranking bias

<table>
<thead>
<tr>
<th>Query</th>
<th>Ranking Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hillary Clinton</td>
<td>0.18</td>
</tr>
<tr>
<td>Bernie Sanders</td>
<td>0.16</td>
</tr>
<tr>
<td>Martin O'Malley</td>
<td>0.07</td>
</tr>
<tr>
<td>Donald Trump</td>
<td>0.10</td>
</tr>
<tr>
<td>Ted Cruz</td>
<td>-0.37</td>
</tr>
<tr>
<td>Marco Rubio</td>
<td>-0.29</td>
</tr>
<tr>
<td>Ben Carson</td>
<td>0.26</td>
</tr>
</tbody>
</table>
Examining ranking bias

<table>
<thead>
<tr>
<th>Query</th>
<th>Ranking Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hillary Clinton</td>
<td>0.18</td>
</tr>
<tr>
<td>Bernie Sanders</td>
<td>0.16</td>
</tr>
<tr>
<td>Martin O'Malley</td>
<td>0.07</td>
</tr>
<tr>
<td>Donald Trump</td>
<td>0.10</td>
</tr>
<tr>
<td>Ted Cruz</td>
<td>-0.37</td>
</tr>
<tr>
<td>Marco Rubio</td>
<td>-0.29</td>
</tr>
<tr>
<td>Ben Carson</td>
<td>0.26</td>
</tr>
</tbody>
</table>

Ranking bias does exist...
Examining ranking bias

<table>
<thead>
<tr>
<th>Query</th>
<th>Ranking Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hillary Clinton</td>
<td>0.18</td>
</tr>
<tr>
<td>Bernie Sanders</td>
<td>0.16</td>
</tr>
<tr>
<td>Martin O'Malley</td>
<td>0.07</td>
</tr>
<tr>
<td>Donald Trump</td>
<td>0.10</td>
</tr>
<tr>
<td>Ted Cruz</td>
<td>-0.37</td>
</tr>
<tr>
<td>Marco Rubio</td>
<td>-0.29</td>
</tr>
<tr>
<td>Ben Carson</td>
<td>0.26</td>
</tr>
</tbody>
</table>

... but no evidence of systemic bias
Does input bias matter?
- Input bias does matter
- Can vary significantly based on the query
  - Even for the same event, different phrasings of queries have widely differing biases

Does the ranking bias exist?
- Yes and varies across queries
- No evidence of systemic bias
Open challenge: How to address search bias?
Open challenge: How to address search bias?

- Modify ranking system to account for bias
- Might lead to reduction in quality of results
Open challenge: How to address search bias?

- Modify ranking system to account for bias
  - Might lead to reduction in quality of results

- Make the bias transparent
  - Keep the current ranking
  - Inform the users about the bias they are seeing
  - Make biases related to query phrasing transparent
Quantifying Search Bias

Ted Cruz is inferred to be republican leaning

John Oliver is inferred to be democrat leaning

https://tinyurl.com/bias-users

@juhi153

CSCW 2017
Quantifying Search Bias

Ted Cruz is inferred to be Republican leaning

John Oliver is inferred to be Democrat leaning

Sean P. Goggins is inferred to be Democrat leaning

https://tinyurl.com/bias-users

@juhi153  Quantifying Search Bias  CSCW 2017