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

Muhammad Bilal Zafar

with Juhi Kulshrestha, Motahhare Eslami, Johnnatan Messias, Saptarshi Ghosh, Krishna Gummadi and Karrie Karahalios
Social media as search platform

- Two-thirds of Americans on social media use it to get news (informational queries)
Social media as search platform

- Two-thirds of Americans on social media use it to get news *(informational queries)*
Social media as search platform

- Two-thirds of Americans on social media use it to get news (informational queries)
Potential bias in search results

Ouch =&gt; Tucker Carlson Obliterates ObamaCare Architect: 'Nobody believes you, doctor'
Potential bias in search results

Steven Crowder
@scrowder

Ouch => Tucker Carlson Obliterates ObamaCare
Architect: 'Nobody believes you, doctor'

Mia Farrow
@MiaFarrow

My daughter in law is a pediatric cardiologist in a
Detroit hosp. Says Obamacare is a lifeline for her patients
Search rankings can shape user opinion

- Users place greater trust in higher ranked items [Pan et al., 2007]

- Biased search results can influence voting patterns [Epstein & Robertson, 2015]

This talk:
Study sources of bias in social media search
This talk: Study sources of bias in social media search

- Identifying sources of search bias
- Quantifying bias in each source
- Studying bias in political searches in Twitter
This talk: Study sources of bias in social media search

- Identifying sources of search bias
- Quantifying bias in each source
- Studying bias in political searches in Twitter
Social media search engine design

Data corpus
Social media search engine design

Data corpus

\[ q \]

\[ i_1, i_2, i_3, \ldots \] (set)

input (relevant items)
Social media search engine design

Data corpus → q → Search engine algo. → ranking

Input (relevant items): i₁, i₂, i₃, ... (set)
Social media search engine design

Data corpus

\[ q \]

\[ \{i_1, i_2, i_3, \ldots\} \] (set)

Search engine algo.

\[ \{i_5, i_1, i_3, \ldots\} \] (ranked list)

input (relevant items)

ranking

output (ranked list)
Output bias may stem from
- Bias in the (input) relevant item set
Social media search engine design

Output bias may stem from
- Bias in the (input) relevant item set
- Bias introduced by the ranking algorithm
This talk: Study sources of bias in social media search

- Identifying sources of search bias
- Quantifying bias in each source
- Studying bias in political searches in Twitter
Quantifying bias in each source

- Quantify bias of each individual item

Diagram:

- Data corpus
- Input (relevant items)
- Search engine algo.
- Ranking
- Output (ranked list)

- Query (q)
Quantifying bias of individual items
Quantifying bias of individual items

- Our method: Infer bias of each individual item from the bias of the author
Quantifying bias of individual items

- Our method: Infer bias of each individual item from the bias of the author

Steve Crowder
@scrowder

Ouch => Tucker Carlson Obliterates ObamaCare Architect: 'Nobody believes you, doctor'
Our method: Infer bias of each individual item from the bias of the author.

"...conservative political commentator..."

Ouch => Tucker Carlson Obliterates ObamaCare Architect: 'Nobody believes you, doctor'
Quantifying bias of individual items

- Our method: Infer bias of each individual item from the bias of the author
- 75% accuracy
- Highly scalable, high coverage
- Other measures can be used [Zafar et al., 2016]
Quantifying bias in inputs (relevant item set)

- Compute bias for each item
- Take the average over the whole set

\[ IB(q) = \frac{\sum_{i=1}^{n} s_i}{n} \]
Quantifying bias in outputs (ranked list)

- 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 in ranking algorithm

Search engine algo.

Ranking bias = Output bias - Input bias
This talk: Study sources of bias in social media search

- Identifying sources of search bias
- Quantifying bias in each source
- Studying bias in political searches in Twitter
• Search queries for
  - 2016 Democratic and Republican debates (*dem debate, #demdebat*, ...)  
  - Presidential candidates (*Hillary Clinton, Donald Trump, ...*)
Key takeaways

- It is not just the algorithm, input data also matters
## High ranking bias

<table>
<thead>
<tr>
<th>Query</th>
<th>Input Bias</th>
<th>Ranking bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hillary Clinton</td>
<td>0.03</td>
<td>0.18</td>
</tr>
<tr>
<td>Ted Cruz</td>
<td>-0.11</td>
<td>-0.37</td>
</tr>
</tbody>
</table>
### High input bias

<table>
<thead>
<tr>
<th>Query</th>
<th>Input Bias</th>
<th>Ranking Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bernie Sanders</td>
<td>0.55</td>
<td>0.16</td>
</tr>
<tr>
<td>Donald Trump</td>
<td>0.19</td>
<td>0.10</td>
</tr>
</tbody>
</table>
Key takeaways

• It is not just the algorithm, input data also matters

• Analyzing the input bias
  – Search topic matters
Effect of search topic

<table>
<thead>
<tr>
<th>Query</th>
<th>Input Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hillary Clinton (D)</td>
<td>0.03</td>
</tr>
<tr>
<td>Bernie Sanders (D)</td>
<td>0.55</td>
</tr>
<tr>
<td>Martin O'Malley (D)</td>
<td>0.64</td>
</tr>
<tr>
<td>Donald Trump (R)</td>
<td>0.19</td>
</tr>
<tr>
<td>Ted Cruz (R)</td>
<td>-0.11</td>
</tr>
<tr>
<td>Marco Rubio (R)</td>
<td>-0.12</td>
</tr>
</tbody>
</table>
### Effect of search topic

<table>
<thead>
<tr>
<th>Query</th>
<th>Input Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hillary Clinton (D)</td>
<td>0.03</td>
</tr>
<tr>
<td>Bernie Sanders (D)</td>
<td>0.55</td>
</tr>
<tr>
<td>Martin O'Malley (D)</td>
<td>0.64</td>
</tr>
<tr>
<td>Donald Trump (R)</td>
<td>0.19</td>
</tr>
<tr>
<td>Ted Cruz (R)</td>
<td>-0.11</td>
</tr>
<tr>
<td>Marco Rubio (R)</td>
<td>-0.12</td>
</tr>
</tbody>
</table>
Key takeaways

• It is not just the algorithm, input data also matters

• Analyzing the input bias
  – Search topic matters
  – Query phrasing can make a big difference
### Query phrasing and search bias

<table>
<thead>
<tr>
<th>Query</th>
<th>Input Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>democratic debate</td>
<td>0.38</td>
</tr>
<tr>
<td>#democraticdebate</td>
<td>0.19</td>
</tr>
<tr>
<td>#demdebate</td>
<td>0.56</td>
</tr>
</tbody>
</table>
Query phrasing and search bias

<table>
<thead>
<tr>
<th>Query</th>
<th>Input Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>democratic debate</td>
<td>0.38</td>
</tr>
<tr>
<td>#democraticdebate</td>
<td>0.19</td>
</tr>
<tr>
<td>#demdebate</td>
<td>0.56</td>
</tr>
</tbody>
</table>
Even for the same topic, query phrasing can greatly affect the input bias.
Query phrasing and search bias

Even for the same topic, query phrasing can greatly affect the input bias.

<table>
<thead>
<tr>
<th>Query</th>
<th>Input Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>republican debate</td>
<td>0.27</td>
</tr>
<tr>
<td>#gopdebate</td>
<td>0.10</td>
</tr>
<tr>
<td>rep debate</td>
<td>0.40</td>
</tr>
</tbody>
</table>
Future work: Addressing search bias

• Alter the ranking algorithm
  – Might lead to reduction in quality of results or lead to unforeseen biases

• Make the bias transparent
  – Keep the current ranking
  – Inform the users about what they are seeing
  – Make biases related to query phrasing transparent
Questions

Paper at:

http://tinyurl.com/se-bias