Measuring Ideological Bias in News Coverage of Political Events by Print Media using Data Analytics

Capstone Project

by

Dev Sheth

(drsp1909@rit.edu)

Dept. of Computer Science,

RIT.
1. **Introduction.**

The role of news organizations in informing public opinion has been evolving as we move into the era of information. Before the internet took over, the primary goal of the fourth estate was to draw public attention to the major events taking place around them and their effects, small and large. Now, however, the public keeps itself informed via social media and a plethora of other communication services. As a result, news reporters increasingly find themselves analyzing events in addition to providing rich content that elaborates on the sparse information dispersed online.

The channels of information distribution have changed as well. Previously, news broadcasting and journalism were primarily sourced through technically skilled journalists, who had been trained in the art of news reporting, in the merits of objectivity, fact-checking, and balanced thoughtful coverage. Today, the information networks that distribute the news do not necessarily follow due process in evaluating the validity of the information. Because of the unique persistence of connectivity of our times, it is far more likely that any news story reaches a person as a social media post by someone they know rather than as a news article by a source known to have journalistic integrity. Indeed, many studies have found that a growing section of the population cites Facebook as its primary news source.

The problem with this is that Facebook has persistently tried to deny the responsibilities of being a media company. This argument by Facebook and other social networks may be true, when evaluating their content distribution networks and algorithms with respect to the traditional definitions of a media company. However, the roles and responsibilities of media companies are not the same as they used to be, and so it is hard to define what type of companies social networks are.
The fact remains that users consume information through social networks, which are fundamentally designed to be networks of family, friends, peers, and generally like-minded individuals across different levels of familiarity. The more familiar a user is to a person in their network, the more they are likely to agree on issues of ideology. Since social networks are designed to offer people a platform to voice their opinions, any story that reaches a user is likely to have passed through a multitude of users who introduce their opinions into the mix. Additionally, because these networks are interconnected, the path any piece of information follows from the source to the user will invariably package it into a form that is more agreeable to the user with each step it takes towards the user’s attention.

This dense network of agreeable connections through which information passes to eventually reach a user is called the information bubble. A user lives within the bubble, unaware of opinions outside the confines of the bubble, because they have actively been filtered out. What is more, the network also makes sure that it echoes the information around the user via multiple sources, repeating the filtered information over and over until it becomes enshrined as fact in the mind of the user. Such a network is said to be an echo chamber.

In order to reach its audiences via social media, any news source has to package the news in a form that is most likely to pass through the networks of agreeable consumers quickly. This becomes easier when the news itself is similar to the audience’s expectations of the same. As a result, the tweaking of the news packaging dissolves into the transformation of the news itself.

This interpretation of news events according to one’s needs has become increasingly popular over the last few years. Under the twin legacies of a devastating recession, multiple wars and stagnating economic growth for lower income households, and that of accelerating growth in technology and the heavy capital investment it attracts, the US population has fundamentally been split into camps
that live in entirely separate realities. Globalization and technological advancements are two wide-scale eventualities that have combined to result in an ideological divide down the boundaries of progress. This, combined with pre-existing cultural differences borne out of changing perceptions of religion and social equalities, has led to both sides being entrenched in their camps, continuously seeking to reinforce the validity of their ideas, and demonize that of “the others.”

The leading national news organizations have taken positions on either side of this divide, mainly driven by the audiences they serve. This has led to an increased emphasis on the accusations of bias in the media. Usually, the bias is said to favor liberal ideologies, according to the right-wing accusers. Unfortunately, the editorial positions of the organizations often tend to creep into the objective reporting of their writers, consciously or otherwise, legitimizing the accusations for some news sources. The process is partly driven by internal biases. But the role of market models is far more significant, since a vast number of the most popular news organizations creating content for online consumption are print newspapers. And for various reasons, print media has gradually turned into loss-leaders for the corporations that own them. It makes economic sense for the media to attempt to drive revenue upwards. To do so, the media must make room for content that its target audience is primed to appreciate, further calcifying the ideological divide in the audience.

A significant amount of research has been carried out in the ideological divides among the audiences of various news outlets. Unfortunately, the same cannot be said for the ideological biases that afflict news media, and the degree to which it can be quantified by inspecting the content it produces.

The goal of this project is to attempt to detect and measure bias in news coverage by the US’s mainstream print media.

Literature Review focuses on some of the research done in this field. Approach outlines some of the possible solutions that can be applied to achieve our goal. Design discusses one such solution that
will be used in this study to devise a model for measuring bias. Results discuss the outcome of the experiment. Finally, Conclusion addresses the implications of the results, and any resultant insight into the problem.
2. Literature Review.

A significant portion of the research concerning ideological biases and news media focuses on the audiences of the media. Baron\textsuperscript{[1]} presents a bias theory that examines both the impact of profit-seeking media and of the impact of journalistic wages on bias. Bernhardt et al.\textsuperscript{[2]} discuss the impact of media bias on electoral decisions. Stone\textsuperscript{[5]} describes a profit-maximization market model where media bias is driven by leveraging the hostile media phenomenon to capture greater market share. Sutter\textsuperscript{[6]} attempts to answer the accusations of liberal media bias using a regression model on newspaper circulation data. Finally, Groseclose and Milyo\textsuperscript{[4]} attempt to measure media bias by comparing the number of citations to political think tanks made by news outlets to those made by members of the Congress. Here, we shall discuss these studies in further detail, before revisiting them when examining the outcomes of this study.

In his paper, Baron\textsuperscript{[1]} offers two market models that would support and explain media bias. The demand-side explanation postulates that audiences have a demand for news stories that are entertaining as well as informative. In order for a news story to be perceived to be entertaining, the story must typically agree with and endorse the reader’s political viewpoint, for the reader to approve of the story and be influenced to consume more stories from the source. This leads to a market-share model where news media outlets look at the competition in the market share and position themselves ideologically so as to capture maximum market share.

The supply-side explanation, on the other hand, revolves around the career benefits to journalists in biasing the stories they write. If it is true that biased stories get more consumption, and are therefore more likely to be prominent in a print newspaper, a journalist may find an incentive to skew the story in that direction. Allowing such creative license to advance their career prospects makes their
job easier, which is directly related to the wages they must be paid to maintain their continued employment with the news organization. If journalists are required to be objective in their writing, the success of their stories depends on the popularity of the stories after publication, and therefore their career growth prospects are not as well-defined, resulting in higher wages. Additionally, the relative ease of the job for a biased reporter lowers the bar for the skills required for a journalist, and therefore increases the talent pool from which hiring can be done, lowering the wages further.

In their paper, Bernhardt et al.\textsuperscript{[2]} discuss the possibility of electoral mistakes arising from the consumption of biased news by a significant proportion of the voters. The paper postulates that since the media can suppress facts but not fabricate them, a biased media source is likely to influence cross-over voting for an election where there is little negative news about the candidates. On the other hand, for elections with extensive focus on the negative news concerning both candidates, strong partisan alliances might be established by following biased news sources and the voting happens along partisan lines.

In his paper, Stone\textsuperscript{[5]} describes how increased emphasis on the hostile media phenomenon primes the market for biased news sources. The paper argues that most consumers prefer news sources that they believe to be unbiased, and yet consume news from sources that align with their ideologies, under the pretext that their source is unbiased. The hostility comes from the assumption by these individuals that the media outlets popular among people with opposing ideologies are the ones that are biased. The market is presumed to be divided into minority left- and right-wing populations and a majority moderate population. The models described for such a market are either a monopolist model, where media outlets hire moderate reporters to maximize moderate voter market share. In contrast, media outlets in the duopoly model hire highly partisan reporters. In such a system,
consumers would still find the partisan reporter on their side of the spectrum to be more moderate than their counterpart, further converting the moderates into partisan voters.

In his paper, Sutter\textsuperscript{[6]} examines the accusations of liberal media bias by conservatives. Sutter\textsuperscript{[6]} attempts to see if liberal bias, if existent, stems from a greater demand for news among liberal consumers. The outcome is mixed, with no evidence of higher news circulation in liberal states, but higher overall news circulation in metropolitan areas.

Groseclose and Milyo\textsuperscript{[4]} devise a method to measure media bias in their paper using ideological scores, which closely resembles the approach of this study. The study counts the number of times a news outlet cites a think tank or a policy group and identifies a Congress member with a similar citation ratio, and compares their Average Adjusted ADA (Americans for Democratic Action) scores\textsuperscript{[4]}. The outcome of the comparison decides the bias score of the media outlet.
3. Data Set.

The goal of this study was to measure ideological bias in news coverage by mainstream national print media outlets. In the interest of studying the most relevant news sources, the first task of the project would be identifying the newspapers that would supply the data set. After reviewing the newspaper circulation figures available online[^3], the following six newspaper outlets were selected:

- USA Today
- The New York Times
- Los Angeles Times
- New York Post
- The Washington Post
- Daily News

Once the sources had been identified, the next step was to decide the time frame for the study. After careful consideration of the time-constraints of the news cycle, it was decided that the period of two weeks in July when the Republican and Democratic National Conventions took place (17 July 2016 – 30 July 2016) would be ideal for this study. The reasoning behind this was that since the national conventions would be an important milestone in the election cycle, they would likely be the most significant period in terms of balanced political news coverage over the duration of the election, except for election night itself.

Once the time range was set, the next step was to source the news stories from the six newspapers during that period. This was performed by using the RIT Library database, [News and Newspapers](#), through which the news stories for each of these newspapers could be accessed in a chronological order. This streamlined the process of locating the news stories.
The next step was to identify which of the stories located in the previous step qualified for the study. It had been decided that since editorials and opinion pieces are inherently biased, the study would utilize only political desk stories that were classified as news, and therefore required the reporter to be objective.

This was, simply put, a monumental task of sorting through over 8000 documents, from which 485 were identified as qualifying for the study, by examining headlines, and content if required.

The next step was reading through and parsing each of these 485 documents to identify the number of bias terms in each document. The bias was categorized into pro-conservative bias, pro-liberal bias, anti-conservative bias, and anti-liberal bias. Once these aggregate terms were generated for each document, the bias ratios for each document were generated. This was the simple percentage ratio of the number of bias terms in a document for a category to the number of total words in that document.

\[
\text{`x' bias ratio} = \frac{\text{# of bias terms in document for category `x'}}{\text{# of words in document}}.
\]

The idea behind this was to scale the bias in documents so that documents with similar categorical biases could cluster together, irrespective of the length of the document.

Eventually, this resulted in a twelve attribute dataset, containing the attributes: sr. no., source no., article no., wordcount, proC, antic, proL, antiL, proC Bias, antic Bias, proL Bias, and antiL Bias, as shown in the figure below.
“Source No.” identified the media outlet the document was sourced from. The range of values is shown below.

### Table 1. Data Sources

<table>
<thead>
<tr>
<th>Source No.</th>
<th>Source Name</th>
<th>Article Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>The New York Times</td>
<td>134</td>
</tr>
<tr>
<td>2</td>
<td>New York Daily News</td>
<td>72</td>
</tr>
<tr>
<td>3</td>
<td>New York Post</td>
<td>63</td>
</tr>
<tr>
<td>4</td>
<td>Los Angeles Times</td>
<td>56</td>
</tr>
<tr>
<td>5</td>
<td>The Washington Post</td>
<td>91</td>
</tr>
<tr>
<td>6</td>
<td>USA Today</td>
<td>69</td>
</tr>
</tbody>
</table>

“Article No.” identified the record ID for each document within a source sub-dataset.

The bias ratios were ratios in the range [0 … 1].
4. System Design

Once the data set was prepared, the next step was identifying an appropriate analytics/clustering technique to be used.

An initial exploration of possible solutions led to studying the feasibility of the IBM Bluemix suite of web services provided by IBM. Bluemix comprises of a list of services that form the basis for IBM products like Watson. After exploring some of these services, both proprietary and third-party, it was decided that the specificity of the services on offer didn’t suit the flexibility required for this particular study. Hence, it was decided that a modified kMeans clustering solution be built in Java for this purpose.

A simplified execution of the kMeans algorithm is given below:

- Initiate $k$ cluster centers.
- Using distance metrics, measure the distances of a data point from each cluster center.
- Assign the data point to the cluster with the nearest cluster center $k$.
- Repeat for all data points.
- For all data points in a cluster, calculate the aggregate distance of a point from every point, to identify the $k$ new cluster centers.
- Repeat process until $k$ centers stabilize.

For our solution, it was clear that kMeans would cluster fastest by initiating the cluster centers to the five natural cluster centers in the data set at the the center and the vertices of a unit-sized 4D object: $(0,0,0,0), (1,0,0,0), (0,1,0,0), (0,0,1,0), \text{ and } (0,0,0,1)$ respectively, denoting the neutral data point, as
well as maximum bias data points, where the values (a, b, c, d) denoted the proC Bias ratio, anti Bias ratio, proL Bias ratio, and antiL Bias ratio respectively.

This would result in the clustering of the data points, which will be described in the following section.

Once the clustering was performed, the cluster distribution for the data points (DP) from a news source was isolated. For each news source, the conservative bias metrics (pro-conservative data points and anti-liberal data points) were aggregated, along with the liberal bias metrics (pro-liberal data points and anti-conservative data points).

The ratio of the absolute difference of these metrics to the total number of data points from the news source gives the bias score for the news source.

\[
\text{Bias Score} = \frac{|(\text{proC DP} + \text{antiL DP}) - (\text{proL DP} + \text{antiC DP})|}{\text{total source DP}}.
\]
5. Results

The kMeans clustering algorithm resulted in the following cluster distribution upon completion.

<table>
<thead>
<tr>
<th>Source Name</th>
<th>Pro-Liberal</th>
<th>Anti-Conservative</th>
<th>Neutral</th>
<th>Anti-Liberal</th>
<th>Pro-Conservative</th>
<th>Bias Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>USA Today</td>
<td>14</td>
<td>11</td>
<td>31</td>
<td>7</td>
<td>6</td>
<td>0.174 L</td>
</tr>
<tr>
<td>The New York Times</td>
<td>17</td>
<td>49</td>
<td>36</td>
<td>19</td>
<td>13</td>
<td>0.254 L</td>
</tr>
<tr>
<td>The Washington Post</td>
<td>21</td>
<td>26</td>
<td>21</td>
<td>17</td>
<td>6</td>
<td>0.264 L</td>
</tr>
<tr>
<td>Los Angeles Times</td>
<td>12</td>
<td>20</td>
<td>15</td>
<td>3</td>
<td>6</td>
<td>0.411 L</td>
</tr>
<tr>
<td>New York Daily News</td>
<td>13</td>
<td>31</td>
<td>15</td>
<td>8</td>
<td>5</td>
<td>0.431 L</td>
</tr>
<tr>
<td>New York Post</td>
<td>6</td>
<td>0</td>
<td>11</td>
<td>21</td>
<td>25</td>
<td>0.635 C</td>
</tr>
<tr>
<td>Total</td>
<td>83</td>
<td>137</td>
<td>129</td>
<td>75</td>
<td>61</td>
<td></td>
</tr>
</tbody>
</table>

As shown in the table, the study finds that USA Today is the least biased news source, which corresponds with its editorial positions, as well as its circulation numbers[^3]. Similarly, the New York Post is the most biased news source, with a strong conservative bias.

A more detailed analysis of the clustering for each source shows that for this election cycle, the news coverage was broadly anti-conservative, which is in line with the Trump candidacy, and his well-documented clashes with various media outlets. Indeed, the anti-conservative coverage alone exceeds the combined coverage for pro-conservative and anti-liberal stories.

Additionally, it is also clear that for both conservative and liberal biases, the election coverage during the conventions was heavily negative and critical of the opponent, instead of being in favor of their respective candidates. This is especially true for liberal sources, where the anti-conservative coverage exceeded the pro-liberal coverage by over 50%.

In yet another interesting insight, The New York Times, well-documented as being liberal, covered more anti-liberal stories than it did pro-liberal stories. In contrast, the New York Post was exclusively pro-conservative, with no anti-conservative stories covered during the convention weeks.
It is clear from the results discussed earlier that most news sources suffer from biased coverage to some extent. It is also apparent that the duopolist market model described by Stone \(^5\) is likely to be in play today, as evidenced by the fact that both The New York Times and The Washington Post are continued to be viewed as exceedingly biased by some conservative voters, where their bias is within relatively normal bounds.

The results also support Sutter’s conclusion \(^6\) that liberal news consumption is high in metropolitan areas, as demonstrated by the moderate to strong liberal bias in five of the six media outlets.

The results, in conjunction with Bernhardt et al.\(^3\), support the fact that a highly negative news cycle splits voters along strictly partisan lines, as demonstrated by the voting numbers from the election. It is evident in the aftermath of the election that with the increased focus on the flaws in the candidates, the ideological divide across party lines has been further entrenched.

---

\(^{15}\)
6. Conclusion

This study proposes a model for measuring bias in media coverage of stories. The outcomes of the study appear to be valid, and in tune with the general perceptions regarding the media, both the outlets in question and the industry as a whole. The evidence of past research appears to endorse the findings of this study as well.

In the absence of a well-defined, standardized, and peer-approved method to measure media bias, it is important to treat the performance of the model as a solitary example. The idea is to support future work in this direction, with the measurement and comparison model potentially being extended to a more crowd-sourced term-classification scheme that eliminates the vulnerabilities of the human error that threatens the objectivity of the model. The model design itself is objective, and therefore useful regardless of the circumstances, as opposed to Groseclose and Milyo’s approach of using the constantly shifting policy positions of elected government officials as a baseline.

Undoubtedly, there is much work to be done before a reliable method to measure media bias is devised. It might yet come to pass that the model proposed by this project doesn’t perform quite as well for other media outlets. But in the absence of an ideal solution, a working solution will do.
7. References


Appendix A

The following figures depict the cluster distribution of the articles by a media outlet.

Fig. A. Cluster Distribution Pie Chart – USA Today

Fig. B. Cluster Distribution Pie Chart – The New York Times
Fig. C. Cluster Distribution Pie Chart – The Washington Post

Fig. D. Cluster Distribution Pie Chart – Los Angeles Times
Fig. E. Cluster Distribution Pie Chart – New York Daily News

Fig. F. Cluster Distribution Pie Chart – New York Post
Appendix B

The following figure illustrates sample terms and phrases that constitute bias in the corresponding contexts.

<table>
<thead>
<tr>
<th>Anti-Conservative</th>
<th>Anti-Liberal</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Isolationist”</td>
<td>“Cocky Democrats”</td>
</tr>
<tr>
<td>“Struggled to commit”</td>
<td>“Leftist darling”</td>
</tr>
<tr>
<td>- Description of Trump’s initial stance on the Benghazi incident, The New York Times</td>
<td>- Description of DNC speaker Elizabeth Warren, New York Post</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pro-Liberal</th>
<th>Pro-Conservative</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Detail-laden policy agenda”</td>
<td>“He got it done ... Beautifully”</td>
</tr>
<tr>
<td>- Description of Hillary Clinton’s election platform, Los Angeles Times.</td>
<td>- Description of Donald Trump’s RNC speech, New York Post</td>
</tr>
<tr>
<td>“Historic role in breaking the gender barrier”</td>
<td>“Get used to it. He’s earned his shot.”</td>
</tr>
<tr>
<td>- Description of the Hillary Clinton’s nomination as the Democratic presidential nominee, Los Angeles Times.</td>
<td>- Description of Donald Trump’s candidacy, New York Post</td>
</tr>
</tbody>
</table>

Fig. G. Sample Bias Terms