Abstract

The increasing prevalence of political bias in news media calls for greater public awareness of it, as well as robust methods for its detection. While prior work in NLP has primarily focused on the lexical bias captured by linguistic attributes such as word choice and syntax, other types of bias stem from the actual content selected for inclusion in the text. In this work, we investigate the effects of informational bias: factual content that can nevertheless be deployed to sway reader opinion.

We first produce a new dataset, BASIL, of 300 news articles annotated with 1,727 bias spans and find evidence that informational bias appears in news articles more frequently than lexical bias. We further study our annotations to observe how informational bias surfaces in news articles by different media outlets. Lastly, a baseline model for informational bias prediction is presented by fine-tuning BERT on our labeled data, indicating the challenges of the task and future directions.

1 Introduction

News media exercises the vast power of swaying public opinion through the way it selects and crafts information (De Vreese, 2004; DellaVigna and Gentzkow, 2010; McCombs and Reynolds, 2009; Perse, 2001; Reynolds and McCombs, 2002). Multiple studies have identified the correlation between the increasing polarization of media and the general population’s political stances (Gentzkow and Shapiro, 2010, 2011; Prior, 2013), underscoring the imperative to understand the nature of news bias and how to accurately detect it.

In the natural language processing community, the study of bias has centered around what we term

---

1 Equal contribution. Lisa Fan focused on annotation schema design and writing. Marshall White focused on data collection and statistical analysis.

2 Dataset can be found at www.ccs.neu.edu/home/luwang/data.html.
made regarding content selection and organization within articles (Gentzkow et al., 2015; Prat and Strömberg, 2013). As shown in Figure 1, though all three articles report on the same event, Huffington Post (HPO) and Fox News (FOX) each frame entities of opposing stances negatively: HPO states an assumed future action of Donald Trump as a fact, and FOX implies Democrats are taking advantage of political turmoil. Such bias can only be revealed by gathering information from a variety of sources or by analyzing how an entity is covered throughout the article.

We define these types of bias as informational bias: sentences or clauses that convey information tangential, speculative, or as background to the main event in order to sway readers’ opinions towards entities in the news. Informational bias often depends on the broader context of an article, such as in the second FOX annotation in Figure 1: gathering new supporters would be benign in an article describing political campaign efforts. The subtlety of informational bias can more easily affect an unsuspecting reader, which presents the necessity of developing novel detection methods.

In order to study the differences between these two types of bias, we first collect and label a dataset, BASIL (Bias Annotation Spans on the Informational Level), of 300 news articles with lexical and informational bias spans. To examine how media sources encode bias differently, the dataset uses 100 triplets of articles, each reporting the same event from three outlets of different ideology. Based on our annotations, we find that all three sources use more informational bias than lexical bias, and informational bias is embedded uniformly across the entire article, while lexical bias is frequently observed at the beginning.

We further explore the challenges in bias detection and benchmark BASIL using rule-based classifiers and the BERT model (Devlin et al., 2019) fine-tuned on our data. Results show that identifying informational bias poses additional difficulty and suggest future directions of encoding contextual knowledge from the full articles as well as reporting by other media.

2 Related Work

Prior work on automatic bias detection based on natural language processing methods primarily deals with finding sentence-level bias and considers linguistic attributes like word polarity (Re-
the document-level, annotators estimate the overall polarities of how the main event and main entities are covered, and rank the triplet’s articles on the ideological spectrum with respect to one another. Before reading the articles, annotators specify their sentiment towards each main entity on a 5 point Likert scale.\(^2\)

On the sentence-level, annotators identify spans of lexical and informational bias by analyzing whether the text tends to affect a reader’s feeling towards one of the main entities. In addition to the main dimension of bias type (lexical or informational), each span is labeled with the target of the bias (a main entity), the bias polarity (positive or negative towards the target), the bias aim towards the main target (direct or indirect), and whether the bias is part of a quote. Bias aim investigates the case where the main entity is indirectly targeted through an intermediary figure (see the HPO example in Figure 1, where the sentiment towards the intermediary entity “Trump Administration” is transferred to the main target, “Donald Trump”). Statistics are presented in Table 1.

**Inter-annotator Agreement (IAA).** Two annotators individually annotate each article triplet before discussing their annotations together to resolve conflicts and agree on “gold-standard” labels. We measure span-level agreement according to Toprak et al. (2010), where we calculate the F1 score of span overlaps between two sets of annotations (details are in the Supplementary). Although the F1 scores of IAA are unsurprisingly low for this highly variable task, the score dramatically increases when agreement is calculated between individual annotations and the gold standard—from 0.34 to 0.70 for informational bias spans and from 0.14 to 0.56 for the sparser lexical spans, demonstrating the effectiveness of resolution discussions.

During the discussions, we noticed several trends that improved the quality of the gold standard annotations. First, the difficulty of being continually vigilant of one’s own implicit bias would sometimes cause annotators to mark policies they disagreed with as negative bias (e.g., a liberal annotator might consider the detail that a politician supports an anti-abortion law as negative bias). Discussions allowed annotators to re-examine the articles from a more neutral perspective. Annotators also disagreed on whether a detail was relevant background or biasing peripheral information. During discussions, they performed comparisons to other articles of the triplet to make a final decision—if another article includes the same information, it is likely relevant to the main event. This strategy reiterates the importance of leveraging different media sources.

For overlapping spans, we find high agreement on the other annotation dimensions, with an average Cohen’s $\kappa$ of 0.84 for polarity and 0.92 for target main entity.

### Table 1: Descriptive statistics of the BASIL dataset

<table>
<thead>
<tr>
<th></th>
<th>NYT</th>
<th>FOX</th>
<th>HPO</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td># Articles</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>300</td>
</tr>
<tr>
<td># Sentences</td>
<td>3,049</td>
<td>2,639</td>
<td>2,296</td>
<td>7,984</td>
</tr>
<tr>
<td># Words</td>
<td>91,818</td>
<td>70,024</td>
<td>62,321</td>
<td>224,163</td>
</tr>
<tr>
<td># Annotations</td>
<td>636</td>
<td>573</td>
<td>518</td>
<td>1,727</td>
</tr>
<tr>
<td>Sentences / Article</td>
<td>30.5 ± 13.8</td>
<td>26.4 ± 10.2</td>
<td>23.0 ± 11.0</td>
<td>26.6 ± 12.2</td>
</tr>
<tr>
<td>Words / Sentence</td>
<td>30.1 ± 14.0</td>
<td>26.5 ± 12.4</td>
<td>27.1 ± 12.5</td>
<td>28.1 ± 13.2</td>
</tr>
<tr>
<td>Annotations / Article</td>
<td>6.4 ± 4.1</td>
<td>5.7 ± 3.8</td>
<td>5.2 ± 3.5</td>
<td>5.8 ± 3.8</td>
</tr>
<tr>
<td>Bias Type</td>
<td>Informational</td>
<td>Lexical</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Informational</td>
<td>468 (73.6%)</td>
<td>421 (73.5%)</td>
<td>360 (69.5%)</td>
<td>1,249 (72.3%)</td>
</tr>
<tr>
<td>Lexical</td>
<td>168 (26.4%)</td>
<td>152 (26.5%)</td>
<td>158 (30.5%)</td>
<td>478 (27.7%)</td>
</tr>
<tr>
<td>Aim</td>
<td>Direct</td>
<td>Indirect</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Direct</td>
<td>574 (90.2%)</td>
<td>485 (84.6%)</td>
<td>462 (89.2%)</td>
<td>1,521 (88.1%)</td>
</tr>
<tr>
<td>Indirect</td>
<td>62 (9.8%)</td>
<td>88 (15.4%)</td>
<td>56 (10.8%)</td>
<td>206 (11.9%)</td>
</tr>
<tr>
<td>Polarity</td>
<td>Positive</td>
<td>Negative</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive</td>
<td>112 (17.6%)</td>
<td>89 (15.5%)</td>
<td>110 (21.2%)</td>
<td>311 (18.0%)</td>
</tr>
<tr>
<td>Negative</td>
<td>524 (82.4%)</td>
<td>484 (84.5%)</td>
<td>408 (78.8%)</td>
<td>1,416 (82.0%)</td>
</tr>
<tr>
<td>Annotations in quotes</td>
<td>205 (32.2%)</td>
<td>299 (52.2%)</td>
<td>217 (41.9%)</td>
<td>721 (41.8%)</td>
</tr>
</tbody>
</table>

Mean and standard deviation shown where applicable. Annotation dimensions show raw counts and their percentage within the dimension in parentheses.

---

\(^2\)The likely effect of annotators’ prior beliefs on their perception of bias will be investigated in future work.

---

4 Media Bias Analysis

### 4.1 Contrasting the Bias Types

**Informational bias outnumbers lexical bias.** As shown in Table 1, the large majority of annotations in BASIL are classified as informational bias. One explanation for its prevalence is that journalists typically make a conscious effort to avoid
biased language, but can still introduce informational bias, either intentionally or through negligence.

For both bias types though, negative bias spans are much more pervasive than positive spans, mirroring the well-established paradigm that news media in general focuses on negative events (Niven, 2001; Patterson, 1996).

**Lexical bias appears early in an article.** We further study differences in characteristics between lexical and informational annotation spans and find that the two bias types diverge in positional distributions. Figure 2 shows that a disproportionate amount of lexical bias is located in the first quartile of articles. A visual inspection indicates that this may be attributed in part to media sources’ attempts to hook readers with inflammatory speech early on (e.g., FOX: “Paul Ryan stood his ground against a barrage of Biden grins, guffaws, snickers and interruptions.”).

In contrast, informational bias is often embedded in context, and therefore can appear at any position in the article. This points to a future direction of bias detection using discourse analysis.

**Quotations introduce informational bias.** We also find that almost half of the informational bias comes from within quotes (48.7%), highlighting a bias strategy where media sources select opinionated quotes as a subtle proxy for their own opinions (see the second HPO and first NYT annotations in Figure 1).

### 4.2 Portrayal of Political Entities

On the document-level, only 17 out of 100 article sets had reversed orderings (i.e. FOX marked as “more liberal” or HPO marked as “more conservative” within a triplet), confirming the ideological leanings identified in previous studies. Here, we utilize BASIL’s span-level annotations to gain a more granular picture of how sources covering the same events control the perception of entities.

Concretely, we examine the polarity of bias spans with target entities of known ideology. As shown in Figure 3, for both bias types, the percentage and volume of negative coverage for liberal entities strongly correspond to the ideological leaning of the news outlet. Note that though NYT appears to have significantly more informational bias spans against conservatives than HPO, this is because NYT tends to have longer articles than the other two sources (see Table 1), and thus naturally results in more annotation spans by raw count.

Moreover, the breakdown of lexical bias distinguishes FOX from the other two outlets: it comparatively has more negative bias spans towards liberals and fewer towards conservatives, even though all three outlets have more conservative entities than liberal ones across the 100 triplets (average of 99.0 conservatives, 72.7 liberals).

### 5 Experiments on Bias Detection

We study the bias prediction problem on BASIL as a binary classification task (i.e., whether or not a sentence contains bias) and as a BIO sequence tagging task (i.e., tagging the bias spans in one sentence at the token-level). We benchmark the performance with rule-based classifiers and the popular BERT model (Devlin et al., 2019) fine-tuned

---

3 The proportion of annotations to article length are similar for all news outlets: one annotation for every 4.1 (for HPO), 4.5 (for FOX), or 4.6 (for NYT) sentences.
on informational and lexical bias spans separately.

**Training Details.** We utilize the pre-trained BERT-Base model and use the “Cased” version to account for named entities, which are important for bias detection. We run BERT on individual sentences\(^4\) and perform stratified 10-fold cross validation. The validation set is used to determine when to stop training and a held out test set is used for the final evaluation of each fold. For the sentence-level classifiers, both our informational and lexical models use 6,819 sentences for training, 758 for validation, and 400 for testing.

Due to the sparsity of our data, we train and test our token-level models only on sentences containing bias spans of the relevant bias type. Our informational and lexical bias sequence taggers use a train/val/test split of 1,043/116/62 sentences and 383/42/23 sentences respectively. Results are shown in Table 2.

**Sentence-level Classifier.** The fine-tuned BERT is better at predicting informational bias than lexical bias, likely because informational bias is better captured by sentence-level context. As a baseline, we select the 4 sentences\(^5\) in each article with the lowest average TF-IDF token scores as containing informational bias. The intuition is that sentences with different content than the rest of the article are more likely to contain extraneous information that the author chose to include to frame the story in a certain way. We find that this simple baseline performs relatively well considering the difficulty of the task, indicating the importance of explicitly modeling context. Future work may consider leveraging context in the entire article or articles on the same story by other media.

**Token-level Classifier.** From Table 2, we see that the BERT lexical sequence tagger produces better recall and F1 than the informational tagger, highlighting the additional difficulty of accurately identifying spans of informational bias. We also use the polarity and subjectivity lexicons from the MPQA website (Wilson et al., 2005; Choi and Wiebe, 2014) as a simple baseline for lexical bias tagging and find that these word-level cues, though widely used in prior sentiment analysis studies, are insufficient to fully capture lexical bias.

In order to evaluate token-level prediction on the larger original test set, we conduct a pipeline experiment with the fine-tuned BERT models where sentences predicted as containing bias by the best sentence-level classifier from cross validation are tagged by the best token-level model. The results reaffirm our hypothesis that while both tasks are extremely difficult, informational bias is more challenging to detect.

### Table 2: Sentence classification (top) and sequence tagging (bottom) results on lexical and informational bias prediction. For the BERT fine-tuning models, the mean from 10-fold cross validation is shown. The minimum standard deviation from cross validation for all BERT models is 3.36, the maximum is 12.44.

<table>
<thead>
<tr>
<th>Token-level</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Lexical Bias</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Polarity lexicon</td>
<td>8.00</td>
<td>0.17</td>
<td>0.33</td>
</tr>
<tr>
<td>Subjectivity lexicon</td>
<td>28.00</td>
<td>0.65</td>
<td>1.28</td>
</tr>
<tr>
<td>BERT fine-tuning</td>
<td>25.60</td>
<td>29.32</td>
<td>25.98</td>
</tr>
<tr>
<td><strong>Informational Bias</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BERT fine-tuning</td>
<td>25.56</td>
<td>14.78</td>
<td>18.71</td>
</tr>
<tr>
<td><strong>Sentence-to-Token pipeline</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lexical bias</td>
<td>12.00</td>
<td>13.64</td>
<td>12.77</td>
</tr>
<tr>
<td>Informational bias</td>
<td>9.52</td>
<td>5.08</td>
<td>6.63</td>
</tr>
</tbody>
</table>

---

\(^4\)BERT’s maximum input length is 512 tokens, which is shorter than most articles in BASIL. We thus treat sentences as passages, rather than using text of fixed length.

\(^5\)BASIL averages 4.1 informational bias spans per article.

6 Conclusion

We presented a novel study on the effects of informational bias in news reporting from three major media outlets of different political ideology. Analysis of our annotated dataset, BASIL, showed the prevalence of informational bias in news articles when compared to lexical bias, and demonstrated BASIL’s utility as a fine-grained indicator of how media outlets cover political figures. An experiment on bias prediction illustrated the importance of context when detecting informational bias and revealed future research directions.

**Acknowledgements**

This research is supported in part by National Science Foundation through Grant IIS-1813341. We thank Philip Resnik, Nick Beauchamp, and Donghee Jo for their valuable suggestions on various aspects of this work. We are also grateful to the anonymous reviewers for their comments.
References


