

# In Plain Sight: Media Bias through the Lens of Factual Reporting

Lisa Fan<sup>1,\*</sup> Marshall White<sup>1,\*</sup> Eva Sharma<sup>1</sup> Ruisi Su<sup>1</sup>  
Prafulla Kumar Choubey<sup>2</sup> Ruihong Huang<sup>2</sup> Lu Wang<sup>1</sup>

<sup>1</sup> Khoury College of Computer Sciences, Northeastern University, Boston, MA 02115

<sup>2</sup> Department of Computer Science and Engineering, Texas A&M University

{fan.lis, white.mars, sharma.ev, su.ruis}@husky.neu.edu  
{prafulla.choubey, huangrh}@tamu.edu, luwang@ccs.neu.edu

## 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<sup>1</sup> 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 stance (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

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

<sup>1</sup>Dataset can be found at [www.ccs.neu.edu/home/luwang/data.html](http://www.ccs.neu.edu/home/luwang/data.html).

<b>Main event:</b> Democratic presidential candidates ask to see full Mueller report
<b>Main targets:</b> Donald Trump, Democratic candidates
<b>HPO:</b> Democrats want access to special counsel Robert Mueller’s investigation into Russian interference in the 2016 presidential election [ <i>before President Donald Trump has a chance to interfere.</i> ] <sub>Trump</sub> ... Sen. Mark Warner said in a statement: [ <i>“Any attempt by the Trump Administration to cover up the results of this investigation into Russia’s attack on our democracy would be unacceptable.”</i> ] <sub>Trump</sub>
<b>FOX:</b> Democratic presidential candidates [ <b>wasted no time</b> ] <sub>Dems</sub> Friday evening demanding the immediate public release of the long-awaited report from Robert S. Mueller III. ... Several candidates, in calling for the swift release of the report, also [ <i>sought to gather new supporters and their email addresses</i> ] <sub>Dems</sub> by putting out [ <b>“petitions”</b> ] <sub>Dems</sub> calling for complete transparency from the Justice Department.
<b>NYT:</b> And on Saturday, one day before Attorney General William Barr released a short summary of Mueller’s findings, former Texas Rep. Beto O’Rourke charged on the campaign trail in South Carolina that you [ <i>“have a president, who in my opinion beyond the shadow of a doubt, sought to, however [<b>ham-handedly</b>]</i> ] <sub>Trump</sub> <i>collude with the Russian government—a foreign power—to undermine and influence our elections.”</i> ] <sub>Trump</sub>

Figure 1: Examples of negative bias from Huffington Post (HPO), Fox News (FOX), and New York Times (NYT) discussing the same event. *Informational bias* and **lexical bias** are highlighted. The target of the bias is noted at the end of each span. Intermediary targets of indirect bias spans are underlined.

**lexical bias:** bias stemming from content realization, or how things are said (Greene and Resnik, 2009; Hube and Fetahu, 2019; Iyyer et al., 2014; Recasens et al., 2013; Yano et al., 2010). Such forms of bias typically do not depend on context outside of the sentence and can be alleviated while maintaining its semantics: polarized words can be removed or replaced, and clauses written in active voice can be rewritten in passive voice.

However, political science researchers find that news bias can also be characterized by decisions

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** (**B**ias **A**nnotation **S**pans on the **I**nformational **L**evel), 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-

casens et al., 2013), partisan phrases (Yano et al., 2010), and verb transitivity (Greene and Resnik, 2009). However, such studies fail to take into consideration biases that depend on a larger context, which is what we try to address in this work.

Our work is also in line with *framing analysis* in social science theory, or the concept of selecting and signifying specific aspects of an event to promote a particular interpretation (Entman, 1993). In fact, informational bias can be considered a specific form of framing where the author intends to influence the reader’s opinion of an entity. The relationship between framing and news is investigated by Card et al. (2015), in which news articles are annotated with framing dimensions like “legality” and “public opinion.” BASIL contains richer information that allows us to study the purpose of “frames,” i.e., how biased content is invoked to support or oppose the issue at hand.

Research in political science has also studied bias induced by the inclusion or omission of certain facts (Entman, 2007; Gentzkow and Shapiro, 2006, 2010; Prat and Strömberg, 2013). However, their definition of bias is typically grounded in how a reader perceives the ideological leaning of the article and news outlet, whereas our informational bias centers around the media’s sentiment towards individual entities. Furthermore, while previous work mostly uses all articles published by a news outlet to estimate their ideology (Budak et al., 2016), we focus on stories of the same events reported by different outlets.

## 3 BASIL Dataset Annotation

Using a combination of algorithmic alignment and manual inspection, we select 100 sets of articles, each set discussing the *same event* from three different news outlets. 10 sets are selected for each year from 2010 to 2019. We use, in order from most conservative to most liberal, Fox News (FOX), New York Times (NYT), and Huffington Post (HPO). **Main events** and **main entities** are manually identified for each article prior to annotation. The political leanings of the main entities (*liberal*, *conservative*, or *neutral*) are also manually annotated. See the Supplementary for details.

**Annotation Process.** To compare how the three media sources discuss a story, annotators treat each article triplet as a single unit without knowing media information. Annotations are conducted on both *document-level* and *sentence-level*. On

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

Table 1: Descriptive statistics of the BASIL dataset. Mean and standard deviation shown where applicable. Annotation dimensions show raw counts and their percentage within the dimension in parentheses.

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.<sup>2</sup>

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 in-

creases 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.

## 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

<sup>2</sup>The likely effect of annotators’ prior beliefs on their perception of bias will be investigated in future work.

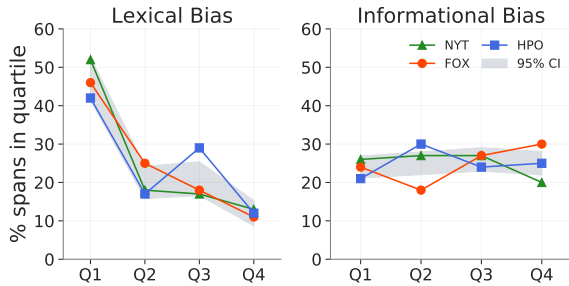


Figure 2: Distribution of lexical and informational bias spans found in each quartile of an article. The shaded area represents the 95% confidence interval for the three outlets combined.

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,

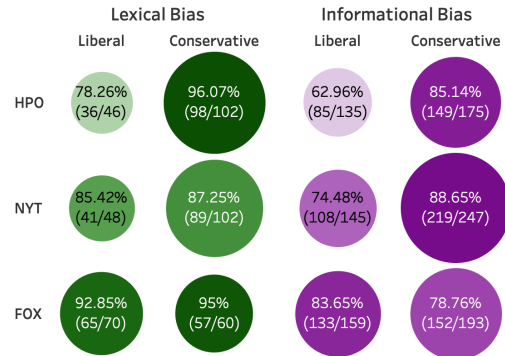


Figure 3: Percentage of bias spans with negative polarity toward targets of known ideology, grouped by media source, bias type, and target’s ideology. For example, in all HPO articles, there are 46 lexical bias spans targeting liberals, 78.26% of which are negative. Larger circle means greater number of spans. Darker color corresponds to higher ratio of negative spans.

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.<sup>3</sup>

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

<sup>3</sup>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.

Sentence-level	Precision	Recall	F1
<i>Lexical Bias</i>			
BERT fine-tuning	29.13	38.57	31.49
<i>Informational Bias</i>			
TF-IDF	25.81	26.23	26.02
BERT fine-tuning	43.87	42.91	43.27

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

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.

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<sup>4</sup> 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<sup>5</sup> in each article with the lowest average TF-IDF token scores as containing

<sup>4</sup>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.

<sup>5</sup>BASIL averages 4.1 informational bias spans per article.

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.

## 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

- Ceren Budak, Sharad Goel, and Justin M Rao. 2016. Fair and balanced? quantifying media bias through crowdsourced content analysis. *Public Opinion Quarterly*, 80(S1):250–271.
- Dallas Card, Amber E Boydston, Justin H Gross, Philip Resnik, and Noah A Smith. 2015. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 438–444.
- Yoonjung Choi and Janyce Wiebe. 2014. +/-effectwordnet: Sense-level lexicon acquisition for opinion inference. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1181–1191.
- Claes De Vreese. 2004. The effects of strategic news on political cynicism, issue evaluations, and policy support: A two-wave experiment. *Mass Communication & Society*, 7(2):191–214.
- Stefano DellaVigna and Matthew Gentzkow. 2010. Persuasion: empirical evidence. *Annu. Rev. Econ.*, 2(1):643–669.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.
- Robert M Entman. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of communication*, 43(4):51–58.
- Robert M Entman. 2007. Framing bias: Media in the distribution of power. *Journal of communication*, 57(1):163–173.
- Matthew Gentzkow and Jesse M Shapiro. 2006. Media bias and reputation. *Journal of political Economy*, 114(2):280–316.
- Matthew Gentzkow and Jesse M Shapiro. 2010. What drives media slant? evidence from us daily newspapers. *Econometrica*, 78(1):35–71.
- Matthew Gentzkow and Jesse M Shapiro. 2011. Ideological segregation online and offline. *The Quarterly Journal of Economics*, 126(4):1799–1839.
- Matthew Gentzkow, Jesse M Shapiro, and Daniel F Stone. 2015. Media bias in the marketplace: Theory. In *Handbook of media economics*, volume 1, pages 623–645. Elsevier.
- Stephan Greene and Philip Resnik. 2009. More than words: Syntactic packaging and implicit sentiment. In *Proceedings of human language technologies: The 2009 annual conference of the north american chapter of the association for computational linguistics*, pages 503–511. Association for Computational Linguistics.
- Christoph Hube and Besnik Fetahu. 2019. Neural based statement classification for biased language. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 195–203. ACM.
- Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. 2014. Political ideology detection using recursive neural networks. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1113–1122.
- Maxwell McCombs and Amy Reynolds. 2009. How the news shapes our civic agenda. In *Media effects*, pages 17–32. Routledge.
- David Niven. 2001. Bias in the news: Partisanship and negativity in media coverage of presidents george bush and bill clinton. *Harvard International Journal of Press/Politics*, 6(3):31–46.
- Thomas E Patterson. 1996. Bad news, bad governance. *The Annals of the American Academy of Political and Social Science*, 546(1):97–108.
- Elizabeth M Perse. 2001. *Media effects and society*. Routledge.
- Andrea Prat and David Strömberg. 2013. The political economy of mass media. *Advances in economics and econometrics*, 2:135.
- Markus Prior. 2013. Media and political polarization. *Annual Review of Political Science*, 16:101–127.
- Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1650–1659.
- Amy Reynolds and Maxwell McCombs. 2002. News influence on our pictures of the world. In *Media effects*, pages 11–28. Routledge.
- Swapna Somasundaran, Josef Ruppenhofer, and Janyce Wiebe. 2008. Discourse level opinion relations: An annotation study. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, pages 129–137.
- Cigdem Toprak, Niklas Jakob, and Iryna Gurevych. 2010. Sentence and expression level annotation of opinions in user-generated discourse. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 575–584. Association for Computational Linguistics.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2-3):165–210.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*.

Tae Yano, Philip Resnik, and Noah A Smith. 2010. Shedding (a thousand points of) light on biased language. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk*, pages 152–158. Association for Computational Linguistics.

## A Sample Annotations

On the right, several sample annotations from the BASIL dataset illustrate some aspects of our annotation schema and highlight characteristics of informational bias.

**Indirect Bias.** Though not as prevalent as bias spans with direct aim, indirect aim is nevertheless important to study because readers may find it more difficult to detect bias consciously when it does not directly implicate the main entity. Indirect bias can be aimed through an intermediary ally or opponent, or may be based on contextual information. In each case, the sentiment towards the intermediary entity alters sentiment toward the main target entity.

Figure 4a shows an example of indirect bias where Donald Trump is negatively targeted via the negative framing of an ally, Donald Trump Jr. Readers are required to know the relationship between the two men in order to notice the bias, and the information itself would be irrelevant to the article were it not for their relationship.

The span from HPO in Figure 4b shows an indirect bias span where contextual information unconnected to the rest of the article reflects negatively on Trump without mentioning him in the text. It requires several leaps in logical thinking: children and families seeking asylum are sympathetic :: turning them away is bad :: Trump wants a border wall :: Trump is framed negatively. This type of informational bias is difficult to detect algorithmically as there is no mention of Trump, the target main entity.

**Informational Bias Strategies.** Inspecting the informational bias spans in our dataset reveals several trends and strategies that journalists tend to

**Main Event:** Trump reverses decision to allow import of elephant trophies  
**Main Entity:** Donald Trump  
**NYT:** On social media, photos were being shared of Mr. Trump’s two elder sons hunting on safari in Zimbabwe, *[including one photo that showed Donald Trump Jr. with a severed elephant tail in one hand and a knife in the other.]Trump*

(a) Indirect negative informational bias against Donald Trump, using the intermediary entity Donald Trump Jr.

**Main Event:** Trump declares national emergency over border wall  
**Main Entity:** Donald Trump  
**HPO:** *[Since 2014, a high proportion of those crossing have been Central American children and families seeking to make humanitarian claims such as asylum.]Trump*  
**FOX:** President Trump said Friday he is declaring a national emergency on the southern border ... *[despite his criticisms of former President Barack Obama for using executive action.]Trump*  
**NYT:** Mr. Trump’s announcement came during a free-wheeling, 50-minute appearance ... *[The president again suggested that he should win the Nobel Peace Prize, and he reviewed which conservative commentators had been supportive of him, while dismissing Ann Coulter, who has not.]Trump*

(b) Example annotations showing negative informational bias from all three media sources for one article triplet.

**Main Event:** Raul Labrador challenges Kevin McCarthy for House majority leadership  
**Main Entities:** Raul Labrador, Kevin McCarthy  
**HPO:** *[Labrador is an ambitious, sometimes savvy politician.]Labrador* He is in Idaho this weekend chairing the state GOP convention.

(c) Example annotation of positive informational bias.

Figure 4: Excerpts showing different types of informational bias, annotated in italics. The target of the negative bias is noted at the end of each span. Underlined entities are intermediary targets in indirect bias spans.

use. The examples from FOX and NYT in Figure 4b show the strategy where objective but tangential information frames the target in a negative light given the context of the article. The example from FOX uses nonessential background information to imply Trump is hypocritical, and the NYT example includes a detail peripheral to the main event that portrays Trump as rambling.

Figure 4c is an example of subtle informational bias where the author’s opinion masquerades as fact. The writing is in a neutral tone and appears objective, but it is actually the author’s perception of the situation and uncovers their bias towards the

topic. The span is categorized as informational bias rather than lexical because there is no way to rephrase or remove parts of the sentence without changing the overall meaning. This span is also an example of the rarer positive bias span.

## B Data Collection

BASIL contains 100 triplets of articles, each with 3 articles about the same main event from the New York Times (NYT), Fox News (FOX), and the Huffington Post (HPO). According to Budak et al. (2016), FOX is considered strongly right leaning, NYT slightly left leaning, and HPO strongly left leaning. As an initial annotation set, 16 triplets of highly visible, polarizing events were directly selected from the media source websites by our annotators.

The remaining triplets were aligned algorithmically from the Common Crawl corpus.<sup>6</sup> Articles with less than 200 words or more than 1,000 words were filtered out, and only political, non-editorial articles published within 3 days of each other were considered. Article similarity was calculated using the cosine similarity of the TF-IDF vectors of each article’s title combined with its first 5 sentences. For each FOX article, the most similar NYT article was found, then the most similar HPO article was found using this pair. An annotator manually selected the final triplets from this list of automatically aligned triplets.

Main event and entities were manually annotated for each article by one annotator. Articles in a triplet share the same main event, which the annotator produced after reading the leads of the three articles. Main entities sometimes differ across the triplet, as stories about the same event can emphasize different characters, but at least one main entity is consistent across each triplet. A single article contains an average of 2.04 main entities and at most five main entities.

During the annotation process, the order of articles is randomized within each triplet and annotators are not aware of the media source of the article. The entire dataset was annotated by three unique annotators.

## C Inter-annotator Agreement

Our study of inter-annotator agreement consists of two parts: the agreement of the text spans selected and the agreement on the dimensions within each

<sup>6</sup><http://commoncrawl.org>

	Exact Matching			Lenient Matching		
	Prec.	Rec.	F1	Prec.	Rec.	F1
<i>Lexical Bias</i>						
A + B	11.04	14.17	12.41	12.34	15.83	13.87
A + C	8.57	9.76	9.13	11.43	13.01	12.17
B + C	–	–	–	15.38	15.38	15.38
<i>Informational Bias</i>						
A + B	19.90	17.22	18.46	39.80	34.33	36.92
A + C	19.47	22.05	20.68	34.40	38.97	36.54
B + C	15.29	10.83	12.68	32.94	23.33	27.32

Table 3: Inter-annotator span agreement for lexical and informational bias. Dashes indicate that there were no exact matching lexical text spans between annotators B and C.

	# Res.	Dimensions (Cohen’s $\kappa$ / % Agr.)		
		Target	Polarity	Aim
A + B	123	0.93 / 93.7	0.84 / 96.3	0.12 / 93.7
A + C	138	0.88 / 89.5	0.75 / 95.0	0.54 / 89.9
B + C	39	0.96 / 96.9	0.92 / 96.9	– / 96.9

Table 4: Number of articles resolved by each annotator pairing, along with Cohen’s  $\kappa$  and percent agreement for IAA on auxiliary dimensions for overlapping spans.

annotation span. To find text span agreement, a similar method to Toprak et al. (2010) is used in which precision, recall, and F1 are calculated between two annotators using the agreement metric from Wiebe et al. (2005), treating one annotator’s spans as the gold standard and the other annotator’s spans as the system. Results are calculated for *exact matching*, where the text spans must overlap exactly to be considered correct, and *lenient matching*, where text spans with any overlaps are considered correct (Somasundaran et al., 2008).

Table 3 shows that span agreement is higher for spans of informational bias than for spans of lexical bias due to the sparsity of lexical bias in our dataset (see Table 1 in the main paper).

Dimension agreement is reported in Table 4 only for lenient matching spans, as the results are not significantly different from that of exact matching spans. Cohen’s  $\kappa$  is used to measure attribute agreement for target, polarity, and aim, and we find high levels of agreement for both polarity and target. Because of the metric’s sensitivity to class imbalance, Cohen’s  $\kappa$  is impractical for measuring the agreement on aim for one annotator pairing (B + C), which had fewer article triplets to resolve and nearly all overlapping lexical annotations were marked as *direct* (31 / 32 spans). To



The screenshot displays a document with several annotations. A blue span highlights the sentence: "Employees would have to dip into their savings or borrow money to pay their bills, as there is no government program to help them." A green span highlights the sentence: "Trump's remarks were riddled with falsehoods on everything from trade to the number of illegal immigrants in the country to the way NATO works." Another green span highlights the phrase "great, great" in the sentence: "During the campaign, he promised a 'great, great' wall made of reinforced concrete that would rise 30 feet or more and extend so..."

The sidebar on the right contains several sections:

- Main Event: Trump continues government shutdown for border wall**
  - How do you think the author feels regarding the main event?
    - Positive
    - Neutral
    - Negative
- Main Entity: Donald Trump**
  - How do you think the author feels regarding this main entity?
    - Positive
    - Neutral
    - Negative
  - How did you feel about this main entity prior to reading the article?
    - Strongly positive
    - Slightly positive
    - Neutral
    - Slightly negative
    - Strongly negative
- Political Stance**
  - Is this article more liberal, neutral or more conservative compared to the other articles in the triple?
    - Note: Within a pair, one should be labeled "More Liberal", one should be labeled "Neutral", and the other should be labeled "More Conservative".
    - More Liberal
    - Neutral
    - More Conservative

Other sections in the sidebar include: **complainig**, **Main Target** (Donald Trump), **Polarity** (Negative), **Aim** (Direct), **Bias Type** (Lexical), **Is the bias in a quote?** (No), and **Additional Notes**.

- (a) Tool with loaded annotations. Informational bias spans are shown in blue, lexical bias spans are shown in green.
- (b) Tool with sentence-level annotation form.
- (c) Document-level annotation form.

Figure 5: Our Javascript annotation tool at various steps.

account for this imbalance, the percent agreement for all attributes is also included in Table 4.

## D Javascript Annotation Tool

A Javascript based tool<sup>7</sup> was developed to annotate our dataset. Annotations created in the tool can be downloaded in JSON format and analyzed or imported at a later date. Users can highlight spans of text or select an entire sentence, then answer dimensional questions (see Figure 5b). Users can also answer document-level questions (see Figure 5c). Figure 5a shows the tool after annotations have been made, where blue spans are informational bias and green spans are lexical bias. In order to alleviate eye strain, annotations of the entire sentence are shown underlined rather than highlighted.

<sup>7</sup><https://github.com/marshallwhiteorg/emnlp19-media-bias>