# Creating a Dataset for Fine-grained Bias Detection in News Articles

Sora LIM<sup>†</sup>, Adam JATOWT<sup>†</sup>, and Masatoshi YOSHIKAWA<sup>†</sup>

† Kyoto University

Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501, Japan

E-mail: †infosora@db.soc.i.kyoto-u.ac.jp, ††adam@dl.kuis.kyoto-u.ac.jp, †††yoshikawa@i.kyoto-u.ac.jp

**Abstract** News articles are nowadays sent out from a variety of news outlets and are published at a very high rate. While many news articles report events in a rather neutral way, unfortunately some exhibit certain bias towards some aspects. As a consequence, The spread of biased news and its consumption by the readers has become a considerable issue. Researchers from multiple domains including social sciences, media studies and professionals from news agencies have made efforts to mitigate this media bias issue. However, due to the lack of publicly available datasets in this field, especially ones containing labels concerning bias on a fine-grained level, it is still challenging to develop methods for effectively identifying bias embedded in new articles. In this paper, we propose a novel news bias dataset which facilitates the development and evaluation of approaches for detecting subtle bias in news articles and for understanding the characteristics of biased sentences. Our dataset consists of 966 sentences from 46 news articles covering 4 different events and contains sentence-level labels obtained by crowd-sourcing. Our dataset can be used for analyzing news bias, as well as for developing and evaluating methods for news bias detection.

Key words News Bias, Crowd-sourcing, Dataset, Media Bias

## 1 Introduction

News articles are not always written in a neutral manner, but may deviate from the norm by using dedicated words, a specific writing style, or a preferred author's viewpoint [7, 12]. Such characteristics of media is referred to as media bias and in the context of news articles as news bias. News bias has been a challenge for a long time in the world of media. Truthfulness, fairness, accuracy, balanced viewpoints have been emphasized in the context of news reporting to avoid news bias, because news can have a large influence on the readers, creating viewpoints and attitudes of people towards social issues, and eventually changing political views and the society [13]. Recognizing news bias is therefore an important goal in the world of media. Nowadays, news articles are mainly published online and are read from various news channels. To monitor and prevent news bias in a timely and efficient manner, first computational approaches [9, 14] have been developed which aim to identify news bias automatically. These methods are based on techniques from natural language processing (NLP) and machine learning. However, detecting news bias automatically is still a major challenge. This can be traced back to several factors.

First of all, news bias is often subtle. Because fairness, factuality, and veracity are considered as crucial for news reporting, the bias often appears with a slight difference of meaning between words and subtle word choice. For instance, it can make a difference when speaking of "climate change" or of "global warming" [16], or when speaking of "illegal immigrants" vs. "undocumented im-

migrants" [11]. As these examples illustrate, identifying such differences, and, thus, media bias overall, typically requires not only applying sentiment analysis on the news articles, but also obtaining deep understanding of reported news events and their context.

Secondly, there is considerable lack of datasets which contain proper labels for news bias. Authors of existing approaches to news bias detection often create their own (manageable) datasets [12]. They use RSS feeds or crawl news websites and then hire experts in the field of linguistics or psychology to annotate the documents. Naturally, creation of these datasets takes a lot of efforts. Moreover, news bias is understood in different ways and labels for bias are provided with different granularity. Particularly, there are very few datasets available that focus on the fine-grained differences on the sentence level. Note that bias in one or just a few sentences in an article might already cause biased opinions in readers and therefore focusing on sentences when evaluating bias is necessary. However, prior datasets mostly indicate the bias status on the document level or on a news source (outlet) level [8, 10]. In the latter case, all articles from the same source receive the same bias label due to the inheritance from their source.

In this research, we create a news bias dataset which can be applied for detecting subtle differences in news articles, and, thus, for analyzing and creating approaches for detecting news bias on a finegrained level. In particular, the dataset contains news bias annotations on the sentence level. Furthermore, unlike other resources, our dataset contains articles about different news topics. In particular, we selected 4 topics covering daily issues on the English news reported between September 2017 and May 2018, which are named as Johnson, Facebook, NFL and North Korea in this paper. Note that the covered news topics are not only associated to politics, but also non-political issues, such as "Republican lawmaker commits suicide amid sexual molestation allegations (named as Johnson)". All these aspects make our dataset unique and promising for future research.

To understand how bias appears in the collected news, we also perform the initial analysis of our dataset. Among other things, we derive statistics concerning the agreement between crowd workers, and compare different user groups w.r.t. the perceived bias.

## 2 Related Work

News bias definitions. Several prior works have focused on media bias in general and news bias in particular. Generally, according to D'Allessio and Allen [4], media bias can be divided into three different types: (1) gatekeeping, (2) coverage and (3) statement bias. Gatekeeping bias is a selection of stories out of the potential stories; coverage bias expresses how much space specific positions receive in media; statement bias, in contrast, denotes how an author's own opinion is woven into a text. Similarly, Alsem et al. [1] divide news bias into ideology and spin. Ideology reflects news outlets' desire to affect readers' opinions in a particular direction. Spin reflects the outlet' s attempt to simply create a memorable story. Given these distinctions, we consider the bias type tackled in this paper as statement bias w.r.t. [4] and as spin bias according to [1].

Hyperpartisan detection datasets. Although news articles are a popular resource for performing research in computational linguistics and natural language processing, the number of datasets dealing with news bias detection is very limited. Noteworthy is, first of all, our preliminary dataset [11]. For this dataset, only a single news event was considered and bias labels were provided on the word level. In contrast, the now proposed dataset covers several events and considers bias on the sentence level. We can further mention the effort to promote the development of novel approaches to media bias detection within the frame of the SemEval 2019 Task 4 "Hyperpartisan News Detection" [10]. For this challenge, the organizers published a large dataset consisting of 754k news articles. However, bias has been defined as hyperpartisan, i.e., w.r.t. a political stance. News articles are then labeled as right, left, and main stream. Furthermore, apart from the relatively small sub-part which was annotated manually, most of the articles were simply labeled according to their news sources. Consequently, the source-level bias annotation has just been automatically used for assigning the document-level bias.

Horne et al. [8] present a large dataset of 136k news articles from 92 news sources for studying the complex media landscape. Both fake sources, satire sources, and hyper-partisan political blogs are considered. Similar to the other datasets, labels are provided on the document level. The dataset of Cremisini et al. [3] is a very recently published dataset containing news articles concerning the Ukrainian Crisis of 2014–2015 from 43 countries. The bias of each article was classified as pro-Russion, pro-Western, or neutral. Although this dataset can be considered as the closest one to our proposal, there are still significant differences. For instance, the authors labeled data on the article level, while our dataset has labels on the level of sentences. Moreover, our dataset covers several news topics and can thus be used to compare biases across different topics and domains.

**Stance classification datasets.** The dataset of Ferreira and Vlachos [5] is an example of a dataset in the area of *stance classification*. Stance classification describes the task of determining the stance of the author of a text document: whether the author illuminates not only one party, but also the opposition. Consequently, stance classification is related to bias detection. However, again, while we consider here subtle differences in the writing, stance classification operates on the document level.

**Fake news detection datasets.** Finally, various datasets have been published for fake news detection. Among the most recent and largest datasets in this regard is the one from Wang et al. [17]. Note that this dataset covers 13,000 manually labeled short statements, but purely in the domain of politics. Also the fake news detection dataset of Perez-Rosas [15] has been published recently. The authors cover seven news domains. Note, however, that fake news detection conceptually differs from bias detection, so that such datasets cannot be used for bias detection research.

### **3** Dataset

#### 31 Constructing News Collection

News bias is always relative (i.e., in relation to explicit or implicit reference) and depends on the context of the news event. One way for handling the relative characteristic, is to compare the content of different news articles which are reporting the same news event, and then, contrast the words used in the articles through the difference news outlets [11]. We also follow this strategy for creating our new dataset. We choose four different news events and collect the news articles reporting those events. The news events used by us are entitled as follows (using the titles of reference articles):

(1) "Republican lawmaker commits suicide amid sexual molestation allegations"  $(\mathbb{E}_1)$ ,

(2) "Trump Clashes with sports world over player protests"  $({}^{(\underline{a}_{2})}$ ,

(3) "Facebook critics want regulation, investigation after data misuse"  $^{(\underline{\ast}3)}$  , and

(4) "Tillerson says U.S. ready to talk to North Korea; Japan wants pressure"  $^{(\pm 4)}$ .

<sup>(</sup>注1):https://reut.rs/2AEnW78

<sup>(</sup>注2):https://www.voanews.com/usa/trump-clashes-sports-world-overplayer-protests

<sup>(</sup>注3):https://reut.rs/2plNZaZ

<sup>(</sup>注4):https://reut.rs/2BfEzFL

Event	News Source	Title of Target Article	Title of Reference Article	
Johnson	Washington Post	Dan Johnson suicide: Lawmaker accused of molesting teen killed himself. His widow calls it a <b>'high-tech lynching</b> .' Dan Johnson, Kentucky lawmaker who killed himself,	(Reuters) Republican lawmaker commits suicide amid sexual molestation allegations	
	The Sun	<ul> <li>claimed he raised woman from the dead</li> <li>Top US politician Dan Johnson kills himself following Face- book post slamming 'judge and jury fake news' after being</li> <li>accused of raping his daughter's 17-year-old pal</li> </ul>		
NFL	ABC News Daily Beast	Trump: 'Standing with locked arms is good, kneeling is not acceptable' Trump's <b>Attack On Black Athletes</b> May Bring a League to Its Feet	(VOA) Trump Clashes With Sports World Over Player Protests	
Facebook	Daily Mail New York Times	Trump-linked Cambridge Analytica tapped 50M Facebook profiles Data Firm Tied to Trump Campaign <b>Talked Business With</b> <b>Russians</b>	(Reuters) Facebook critics want regulation, investigation after data misuse	
North Korea	Newsweek ABC News	U.S. Will Talk To North Korea <b>'Until The First Bomb</b> <b>Drops,'</b> Rex Tillerson Says Tillerson tries to quell anxieties at State Dept. amid ques- tions about his future	(Reuters) Tillerson says U.S. ready to talk to North Korea; Japan wants pressure	

Table 1 Sample news articles in our dataset

These events are in connection with allegation of sexual molestation, conflict between racism and patriotism, a data firm intervened election, and international relationships, respectively. In the following, we refer to these events shortly as (1) JOHNSON, (2) NFL, (3) FACEBOOK, and (4) NORTH KOREA. Table 1 shows brief examples of each news event.

For collecting news articles on the same event from different news outlets, we choose Google News which provides clustered news articles on the main issues of the current time. We collect the articles linked from the Google news' pages and extract the title and text content (after manual inspection in order to exclude image-only or unrelated articles). The resulting dataset consists of 44 news articles for the event JOHNSON, 103 articles for the event FACEBOOK, 371 articles for NFK, and 39 articles for NORTH KOREA. The titles of the articles belonging to each news event are very diverse, indicating that the news articles emphasize different aspects and therefore having potentially different degrees of bias. For example, for the event JOHNSON, one news article emphasizes the suicide event as "hightech lynching", while another one degrades the event by referring "claimed he raised woman from the dead". Also the word choice (e.g., "raping" vs. "sexual molestation") differs significantly. As our goal is to annotate the sentences, we split all the news articles into sentences by using the widely used module from NLTK [2].

#### 32 Annotation of News Bias

To overcome scalability issue concerning annotations, crowdsourcing has been recently widely used [18]. We also rely on this approach in our research to obtain labels for the news article sentences w.r.t. their degree of bias. Crowdsourcing has also the additional benefit of allowing for understanding bias from the point of view of ordinary users who are likely to be typical readers of news articles. We use Figure  $Eight^{(\pm 5)}$  as an underlying crowd-sourcing platform.

We note that it is rather difficult to provide bias-related labels such as binary judgements on each sentence of news articles, as the bias may depend in various ways on the news event and its context. Thus, to effectively design the bias labeling task, we pick one news article as a *reference article* for each analyzed news event. The reference article has been selected from well-known news agencies which supply news items to the news outlets like Reuters or AP. Typically, such news agencies are known to be neutral and containing least bias. Having a reference article serves two functions: Firstly, reading first the reference news article about the described news event lets annotators become first familiar with the overall news event. If no such reference article were provided, crowdworkers might miss important context information to properly judge the bias of sentences later on. Secondly, assuming that the reference article is relatively bias-free, it might be easier for crowd-workers to recognize biased sentences in the target articles (e.g., due to noticing different word choices, e.g., "raping" vs. "sexual molestation").

As the preliminary stage to construct a large news bias dataset, we first run the task with a subset of articles which are equally sampled according to their news outlet's political stance instead of labeling whole news articles in the dataset. For the sampling step, we make use of Media Bias/Fact Check<sup>( $\pm 6$ )</sup> website which provides the information of political position of over 2900 media sources according to the ideology spectrum. Specifically, for each news event, we picked

<sup>(</sup>注5):https://www.figure-eight.com

<sup>(</sup>注6):https://mediabiasfactcheck.com

Name	Description	Content		
event	news event of the articles	{Donjonson, Facebook, NFL, North Korea}		
date_event	published date of the news articles	{2017-12-15, 2018-03-18, 2017-09-24, 2017-12-13}		
id_article	individual id of the news article	from 0 to 46		
loc_sentence	position of the sentence in the news articles	title, from 1 to 20		
source	news sources of the articles	{Washington Post, New York Times,}		
source_bias	political stance of the news source	(laft laft contan loost right contan right)		
	of the article	{ien, ien-center, ieasi, ingin-center, ingin}		
ref	the reference article	Reuters, Voice of America (NFL)		
url	urls of the news article			
article_bias	labeled bias degree of the target article	{neutral, slightly biased but acceptable, biased, very biased}		
sentence_bias	labeled bias degree of each sentences	{neutral, slightly biased but acceptable, biased, very biased}		
preknow	rater's prior knowledge of the news event	{yes, no}		

Table 2	Information	stored	in (	our c	lataset

Sentence	Label		
In the post, written on Wednesday, Johnson paid tribute to his family, saying he had suffered post-traumatic stress disorder for 16 years -			
"a sickness that will take my life".			
We've been promised by Google, Facebook, and other social sites that our personal information is protected and that when some of our			
information is provided to third parties, our identity will never be made known.			
Secretary of State Rex Tillerson is ready to talk about talking to North Korea.			
This Wednesday, Kentucky state representative Dan Johnson was found dead by a bridge in Mount Washington with a "single gunshot			
wound" to the head, according to Bullitt County Coroner Dave Billings.			
Table 2 Example contanges from our dataset with the labels obtained via around sourcing task			

Table 3 Example sentences from our dataset with the labels obtained via crowd-sourcing task

for each political position randomly four news articles.

In the annotation tasks, crowd-workers were asked to rate the degree of bias for each news sentence in each target article based on the four-scale category: neutral and not biased, slightly biased, biased, and very biased. After the completion of the annotation for sentences in each target news article, the crowd-workers are also asked about the bias degree of the entire target article compared to the reference article, and whether they knew about the reported news event before. In total 28 users contributed to our labeling task after excluding untrustworthy annotators. Overall, we collected 4,515 annotations for 966 sentences (5 annotations per each sentence) from 46 news articles through 215 unit labeling tasks. An overview of our dataset is given in Table 2.

#### 4 Analysis of Bias Annotation

As we let five different crowd-workers rate the bias degree of 21 sentences (the article's title and the first 20 article's sentences) in a job unit, 996 sentences are tagged with multiple bias category, resulting in 29.97% neutral and not biased, 34.37% slightly biased but acceptable, 28.17% biased, and 7.49% very biased – as shown in Figure 1. When we converted bias degrees to binary judgments (such as the first two considered as unbiased and the latter two as biased), 35.55% sentences were assigned to biased sentence category. Table 3 shows examples of the users' ratings.

After we collect the bias annotations, we examined the inter-rater reliability among the five crowd-workers' answers. We calculated Fleiss' kappa score [6], which is widely used measure to check the extend of the answer agreement among any number or raters giving categorical ratings to a fixed number of items (in our case, five raters giving 4-category ratings to 21 items). The mean scores calculated over all the target articles in each news events are -0.062, -0.078, -0.014 and -0.049 for the labels concerning the news events JOHNSON, FACEBOOK, NFL, and NORTH KOREA, respectively. As shown in Figure 2, the agreement tendency is relatively low. We suspect this low agreement be the result of different sensitivity to the bias between users, as also reported during the creation of a similar dataset in [11].

During the bias annotation, we also asked the raters whether they already had any knowledge about the reported news events. Based on the answer to this question, we categorized the crowd-workers into the user group "people who knew" and "people who didn't know" the target news event. Here our hypothesis was that when people have some knowledge about the news event, they may already have formed their own opinion based on the previously acquired information from the news media (which was uncontrolled for us as task providers). Consequently, the hypothesis was that this user group might differ more from the "people who didn't know" group w.r.t. bias. Figure 3 shows the distribution of the user groups according to the prior knowledge about the news events. To determine these user-group differences, we analyzed the inter-rater agreement scores according to the two user groups. The results showed that the average agreement scores differ between these two



Figure 1 Bias label distribution on document-level



Figure 2 Inter-rater reliability on the Crowdsourcing result: four-scale case



Figure 3 Previous knowledge about the news events

user groups by indicating a higher agreement score in the "people who knew" group. However, as Figure 3 shows, overall there was a considerably small ratio of people who knew the news event.

#### **5** Conclusions

Detecting news bias is a challenging task for computer science as well as linguistics and media research areas due to the subtle nature and heterogeneous, diverse kinds of biases. In this paper, we presented a corpus of news articles where sentences of the news articles have been labeled in a crowd-sourcing task concerning their degree of bias. To annotate subtle bias by using ordinary people on crowd-sourcing who can be considered as general news readership, we provided a standard of comparison of the news content as the reference article. We then analyzed these annotations in regards to perceived bias. Based on our analysis, we conclude that the prior knowledge about news event can be a major factor for bias annotation.

#### References

[1] Karel Jan Alsem, Steven Brakman, Lex Hoogduin, and Gerard Ku-

per. The impact of newspapers on consumer confidence: does spin bias exist? *Applied Economics*, 40(5):531–539, 2008.

- [2] S Bird, E Loper, and E Klein. Natural language processing with python o' reilly media inc. 2009.
- [3] Andres Cremisini, Daniela Aguilar, and Mark A. Finlayson. A Challenging Dataset for Bias Detection: The Case of the Crisis in the Ukraine. In *Proceedings of the 12 International Conference on Social, Cultural, and Behavioral Modeling*, SBP-BRiMS'19, pages 173–183, 2019.
- [4] Dave D'Alessio and Mike Allen. Media bias in presidential elections: A meta-analysis. *Journal of communication*, 50(4):133–156, 2000.
- [5] William Ferreira and Andreas Vlachos. Emergent: a novel dataset for stance classification. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT'16, pages 1163–1168, 2016.
- [6] Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.
- [7] Robert A Hackett. Decline of a paradigm? bias and objectivity in news media studies. *Critical Studies in Media Communication*, 1(3):229–259, 1984.
- [8] Benjamin D. Horne, Sara Khedr, and Sibel Adali. Sampling the News Producers: A Large News and Feature Data Set for the Study of the Complex Media Landscape. In *Proceedings of the Twelfth International Conference on Web and Social Media*, ICWSM'18, pages 518–527, 2018.
- [9] C Hutto, Dennis Folds, and Darren Appling. Computationally detecting and quantifying the degree of bias in sentence-level text of news stories. In *Proceedings of Second International Conference on Human and Social Analytics*, 2015.
- [10] Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. SemEval-2019 Task 4: Hyperpartisan News Detection. In Proceedings of The 13th International Workshop on Semantic Evaluation (SemEval 2019). Association for Computational Linguistics, 2019.
- [11] Sora Lim, Adam Jatowt, and Masatoshi Yoshikawa. Understanding characteristics of biased sentences in news articles. In Proceedings of the CIKM 2018 Workshops co-located with 27th ACM International Conference on Information and Knowledge Management (CIKM 2018), Torino, Italy, October 22, 2018, 2018.
- [12] Fred Morstatter, Liang Wu, Uraz Yavanoglu, Stephen R Corman, and Huan Liu. Identifying framing bias in online news. ACM Transactions on Social Computing, 1(2):5, 2018.
- [13] David Niven. *Tilt?: The search for media bias*. Greenwood Publishing Group, 2002.
- [14] Souneil Park, Seungwoo Kang, Sangyoung Chung, and Junehwa Song. NewsCube: delivering multiple aspects of news to mitigate media bias. In *Proc. of SIGCHI on Human Factors in Computing Systems*, pages 443–452, 2009.
- [15] Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. Automatic detection of fake news. In *Proceedings* of the 27th International Conference on Computational Linguistics, COLING'18, pages 3391–3401, 2018.
- [16] Jonathon P Schuldt, Sara H Konrath, and Norbert Schwarz. "global warming" or "climate change"? whether the planet is warming depends on question wording. *Public Opinion Quarterly*, 75(1):115– 124, 2011.
- [17] William Yang Wang. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, ACL'17, pages 422–426, 2017.
- [18] Arkaitz Zubiaga, Maria Liakata, Rob Procter, Kalina Bontcheva, and Peter Tolmie. Crowdsourcing the annotation of rumourous conversations in social media. In *Proc. of WWW 2015*, pages 347–353, 2015.