A Multidimensional Dataset Based on Crowdsourcing for Analyzing and Detecting News Bias

Michael Färber
Karlsruhe Institute of Technology (KIT)
Karlsruhe, Germany
michael.faerber@kit.edu

Victoria Burkard
Karlsruhe Institute of Technology (KIT)
Karlsruhe, Germany
ujepj@student.kit.edu

Adam Jatowt
Kyoto University
Kyoto, Japan
adam@i.kyoto-u.ac.jp

Sora Lim
Kyoto University
Kyoto, Japan
infosky.sora@gmail.com

ABSTRACT

The automatic detection of bias in news articles can have a high impact on society because undiscovered news bias may influence the political opinions, social views, and emotional feelings of readers. While various analyses and approaches to news bias detection have been proposed, large data sets with rich bias annotations on a fine-grained level are still missing. In this paper, we firstly aggregate the aspects of news bias in related works by proposing a new annotation schema for labeling news bias. This schema covers the *overall bias*, as well as the bias dimensions (1) *hidden assumptions*, (2) *subjectivity*, and (3) *representation tendencies*. Secondly, we propose a methodology based on crowdsourcing for obtaining a large data set for news bias analysis and identification. We then use our methodology to create a data set consisting of more than 2,000 sentences annotated with 43,000 bias and bias dimension labels. Thirdly, we perform an in-depth analysis of the collected data. We show that the annotation task is difficult with respect to bias and specific bias dimensions. While crowdworkers' labels of *representation tendencies* correlate with experts' *bias* labels for articles, *subjectivity* and *hidden assumptions* do not correlate with experts' *bias* labels and, thus, seem to be less relevant when creating data sets with crowdworkers. The experts' article labels better match the inferred crowdworkers' *article* labels than the crowdworkers' *sentence* labels. The crowdworkers' countries of origin seem to affect their judgements. In our study, non-Western crowdworkers tend to annotate more bias either directly or in the form of bias dimensions (e.g., *subjectivity*) than Western crowdworkers do.

KEYWORDS

Media Bias, Crowdsourcing, Text Mining, News Articles

ACM Reference Format:
Michael Färber, Victoria Burkard, Adam Jatowt, and Sora Lim. 2020. A Multidimensional Dataset Based on Crowdsourcing for Analyzing and Detecting News Bias. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20), October 19–23, 2020, Virtual Event, Ireland.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3340531.3412876

CIKM '20, October 19–23, 2020, Virtual Event, Ireland

ACM ISBN 978-1-4503-6859-9/20/10.
https://doi.org/10.1145/3340531.3412876

1 INTRODUCTION

Media bias detection [8] has recently gathered much attention in research communities along with other related tasks, such as fake news detection, rumor detection, and satire detection. Specifically, detecting media bias in news articles is of great importance because news articles remain the primary source to acquire information and to form opinions about events [8]. News bias detection systems can, for instance, be combined with browser plug-ins to assist online news readers with awareness of biased texts [4]. Also, a supportive system integrated in journalistic processes can enable journalists to receive immediate feedback on bias as they are writing [17].

As automated approaches to news bias detection need to be evaluated and often trained, reliable data sets with ground truth annotations (i.e., labeled news articles) are of great importance. Several data sets for fake news and media bias detection exist [6, 19, 20]. However, due to differences in the task setup and labeling, data sets for *fake news* detection (e.g., [22]) are not applicable for *media bias* detection. Furthermore, existing data sets for media bias analysis and detection have the following limitations: (1) Although a few data sets were created by experts based on an elaborate process [6], their comparably small size often limits their use. (2) Other data sets, such as the one from the SemEval-2019 Task 4 [10], are larger in size, but have rather coarse level labels, with most articles labeled according to their news source purely. (3) Remaining data sets for media bias detection were created by crowdsourcing, but are still considerably small in size and use different bias definitions (e.g., having a reference news article as reference point for judging bias [12] or focusing on the sentiment aspect of bias [1, 26]).

In this paper, we aim to create a large-scale benchmark data set for news bias detection. To this end, we firstly distinguish – next to the *overall bias* – three key aspects of bias (called *bias dimensions* in the following) based on a literature review: *hidden assumptions*, *subjectivity* and *representation tendencies*. We use these dimensions as targets for labeling, as we believe that bias is a multidimensional phenomenon that can potentially be modeled more suitably by its different aspects. Thus, we want to study whether a direct labeling of bias is more difficult for non-expert users, particularly on

Table 1: Overview of data sets for news bias detection and analysis.

Name	Content	Labels Related to Bias	Labels Provided by
Ukraine Crisis Articles [6]	4,538 news articles	Bias: "Pro-Russian," "Pro-West," "Neutral"	Experts
SemEval 2019 [10]	750,645 news articles	Article bias, overall publisher bias	Experts, Crowdworkers
NewsWCL50 [9]	50 news articles	Target concepts, frame properties	Experts
Starbucks [11]	1,235 news sentences	Bias w.r.t. a reference article	Crowdworkers
NFNJ [12]	966 news sentences	Bias w.r.t. a reference article	Crowdworkers
Ukraine Crisis Sentences (ours)	2,057 news sentences	Bias, subjectivity, hidden assumptions, representation tendencies	Crowdworkers

sentence level, than labeling specific bias dimensions, and whether bias can be detected by looking at associated psychological aspects showing up in linguistic characteristics of biased news articles.

Secondly, we present a new data set generation approach using crowdsourcing. Our main goal is to establish a labeling procedure for data which does not rely on experts alone, but rather a scalable crowdsourcing-based solution. In our case, bias is always with respect to a particular direction or target, rather than being undirected *bias* as in [11, 12].

Thirdly, we use our approach to label a large data set comprised of labels with respect to bias on an article as well as sentence level. The data set covers **2,057 sentences** from 90 news articles published in 33 countries. All articles deal with the Ukraine crisis and were selected from the data set of Cremisini *et al.* [6]. We extend this data set by **44,547 labels** in total (43,197 sentence labels and 1,350 article labels). Our data set is provided to the research community online.[1]

Lastly, we analyze the created data set. For instance, we calculate the correlations between the crowdworkers' annotations and the expert's annotations. We also shed light on the differences between labels on sentence and article level, as well as with regard to the crowdworkers' background. Overall, our data analysis shows that the data set allows us to analyze several news bias aspects for the first time.

Overall, our main contributions can be summarized as follows:

- We propose an annotation schema for news bias detection.
- We present a scalable approach for data set generation using crowdsourcing and the proposed bias annotation schema.
- We show the feasibility of our approach by constructing a news bias data set consisting of 2,057 sentences from 90 articles and provide this data set to the public.
- Given our data set, we analyze the crowdworkers' perceptions of bias and the single bias dimensions.

The rest of the paper is structured as follows: In Section 2, we outline existing data sets and approaches for news bias detection, while in Section 3, we describe the bias dimensions used in this work. In Section 4, we present our approach for large-scale news bias data set generation. Section 5 is dedicated to the in-depth analysis of the generated data set, while Section 6 discusses the impact and use cases of this work. We conclude the paper in Section 7.

[1]See **https://doi.org/10.5281/zenodo.3885351** and our repository https://github.com/michaelfaerber/ukraine-news-bias containing the source code used for Section 5.

2 RELATED WORK

In this section, we first outline currently available data sets on news bias and then present approaches to news bias detection.

Data Sets for News Bias Detection. Published data for media bias detection and especially news bias detection are diverse in that they focus on different aspects indicating bias. An overview summarizing the most important data sets' characteristics is given in Table 1. The existing data sets were labeled only according to the occurrence of *bias* itself, which is – as being a complex and abstract concept – inherently difficult to be recognized. Furthermore, bias was usually not recognized in direction to any named target (e.g., Russia) [11, 12], which increases ambiguity and complexity of the annotation task. Another difference is that the previous data sets were also annotated mainly on a coarse level, either on the level of news sources [10] or on the level of entire articles [6, 9].

Crowdsourcing for Media Bias Analysis. Several publications deal with media analysis using crowdsourcing. Budak *et al.* [5] revealed political slants in news portals. Here, crowdworkers label party members with respect to bias toward a certain party. Pennycook and Rand [18] focused on media perception and showed high correlations between laypeople's assessments of trustworthiness and professional fact checkers at the news outlet level. Benoit *et al.* [3] presented an approach in the field of stance detection using a crowdsourced data set with sentence-labeled party manifestos. Park *et al.* [16] captured slants of news articles via the social news website *NewsCube2.0*. Lim *et al.* [11] analyzed bias on sentence and word level in news articles to reveal linguistic features, such as negative subjects, with the help of crowdworkers.

All these outlined papers comparing annotations by laypeople and experts suggest or presuppose a high correlation between crowdsourced and expert-labeled data. They state that a layperson can, in fact, replace experts up to a certain level of data quality. They also agree achieving a high inter-annotator agreement is one of the most challenging factors with regard to crowdsourcing.

Approaches to News Bias Detection. In the last ten years, only a few computational approaches were published that derive a bias score based on news article features. They can be distinguished in several ways and are categorized in the following.

A common approach is to handle biased news texts as text categories. For instance, Recasens *et al.* [21] trained a classifier based on Wikipedia edits. Moreover, they performed a linguistic analysis to identify word groups indicating bias. All approaches that use classifiers for detecting certain news bias indirectly adopt the bias definition that was chosen within the data annotation. In the case of biased news, those approaches output a bias score that indicates the probability of a given article to be biased.

Another group of approaches handles news bias on a deeper level and examines articles with regard to the goals of biased news texts: Authors aim to arouse feelings to shape one's opinions and mind or to result in actions. Hence, some approaches inspect the underlying tone of an article and rate sentiment by extracting sentiment words [1, 26]. Another aspect is to find an underlying opinion by evaluating the representation of the "target of bias," which are, for instance, persons or abstract concepts. In this context, Ogawa *et al.* [15] referred to their technique as "stakeholder mining," as they examined important entities as stakeholders in the text. Hamborg *et al.* [9] extracted framing properties as well as target words/phrases concerning the subjects of framing from political news articles.

3 BIAS DIMENSIONS OF THE DATA SET

We argue that the specific aspects of bias can be traced to the following *bias dimensions*:[2]

(1) Hidden Assumptions and Premises. Recasens *et al.* [21] found that specific verbs ("assertive," "factive") and word groups ("hedges") are strongly related to bias, as they either slightly doubt, diminish, or increase the informative power of a statement. For example, "he revealed (...)" versus "he stated (...)" increases the informative power of the subsidiary sentence. Beyond that, logical fallacies are an important credibility indicator [25]. For instance, an "appeal to fear fallacy" occurs when the author excessively stresses possible dangers to increase fear resulting in more support of an alternative. Logical fallacies are difficult to recognize even by experts [25]. For this bias dimension, we take the aforementioned word groups and logical fallacies to form a new and simplified category called *hidden assumptions and premises*. It includes unjustified statements presented as generally accepted while different views remain unspoken.

(2) Subjectivity. Biased news articles show subjectivity as distinctive characteristic as revealed by Nakashole and Mitchell [14]. Subjective sentences are opinionated and judgemental. Since most news articles are not written from the first-person perspective, subjectivity in articles is not expressed in a direct manner. Instead, subjective word choices are used, such as "terrorist groups" versus "paramilitary groups". They are classified as "one-sided terms" and "subjective intensifiers" [21].

(3) Framing. Entman [7] defines framing as selecting "some aspects of a perceived reality and making them more salient in a communicating text, in such a way as to promote a particular problem definition, causal interpretation, moral evaluation, and/or treatment recommendation for the item described." Several approaches tackle news bias by identifying frames or connected attributes [2, 13]. We introduce the bias dimension of *framing* to refer to the evaluation of targets of bias in news articles. In order to simplify the framing annotation for crowdworkers, possible *targets* of bias and *evaluations* are given. In the context of the Ukraine crisis, the governments of Russia, Ukraine, and the West are considered as *bias targets* and possible *evaluations* are *positive*, *negative*, and *neutral*.

[2] *Sentiment* is not considered by us, since sentiment labels can be obtained to a high degree automatically nowadays, as well as other related data sets already contain it. Furthermore, sentiment is partially embraced in the framing dimension we use.

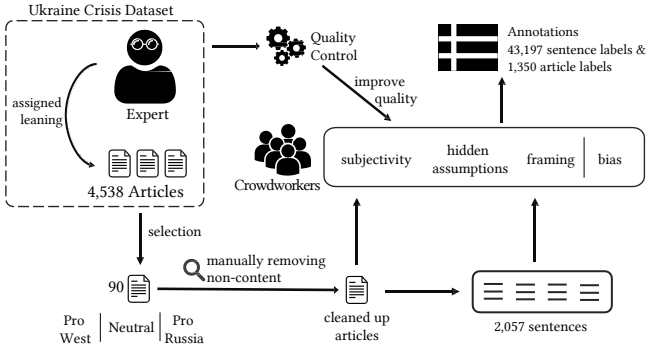

Figure 1: Overview of our proposed bias annotation process.

Overall Bias. We complement our bias annotation schema with *bias itself*. Note that in our schema, bias is always oriented toward a specific target. For instance, in the context of the Ukraine crisis, Cremisini *et al.* [6] used *Pro-West* and *Pro-Russia* as the two possible tendencies regarding bias. We follow Cremisini *et al.* and use the same tendencies for bias annotations on the sentence level.

4 CROWDSOURCING NEWS BIAS DATA SETS

Our approach to bias data set generation is depicted in Figure 1. We first obtain appropriate news articles for the annotation and pre-process them for crowdsourcing (Sec. 4.1). Crowdworkers then rate the article sentences with respect to each bias dimension, as well as bias itself (Sec. 4.2.1). Finally, we calculate the biases on the article and sentence levels in a post-processing step (Sec. 4.3).

4.1 Article Collection

4.1.1 Collection Procedure. The input of the annotation process are news articles. Without loss of generality, we select the data set of Cremisini *et al.* [6] that contains URLs of news articles related to the Ukraine crisis to retrieve the news articles and split each article into sentences. We choose this data set for several reasons: (1) As it contains bias labels on the article level provided by an expert, we can later compare the bias labels of crowdworkers on the sentence level and (inferred) on the article level with expert labels; (2) We can reuse the assigned leanings for our selected articles; (3) Since the data set deals with a recent and relevant topic, we can assume that most crowdworkers are familiar with it.

4.1.2 Collection Statistics. To have an equal distribution, we chose 30 articles of each leaning, for a total of 90 articles (i.e., 30 pro-West, 30 neutral, and 30 pro-Russia). Each article contains, on average, 23 sentences, resulting in *2,057 sentences* to be annotated. The country of origin is evenly distributed for each leaning. Unfortunately, pro-Russian articles in the data set of Cremisini *et al.* [6] all originated from Russia. Thus, we vary these only by article publisher. The maximum length of each article is restricted to 80 sentences, plus title and header sentence.

4.2 Crowdworkers' Annotation

We used the crowdsourcing platform *Appen* (https://www.appen.com), formerly known as Figure Eight and CrowdFlower. To ensure

Table 2: Example sentences with labels.

Sentence	Label (avg./maj./intens.)
"Order in the country can only be restored through dialogue and democratic procedures, rather than with the use of armed force, tanks and aircraft."	Hidden assumpt.: 0.0/ 0.0/ 0.0 Subjectivity.: 1.0/ 0.0 /2.5 Framing Russia: 0.0/ 0.0/ 0.0 Framing Ukraine: -0.8/ 0.0/ -2.0 Framing West: 0.0/ 0.0/ 0.0 Bias Pro-Russia: 1.0/ 1.0/ 1.0 Bias Pro-West: 0.8/ 1.0/ 1.0
"High Time to Resolve Ukrainian Crisis - Kazakh President Lavrov added that the Ukrainian crisis as well as 'the Western anti-Russia campaign including the illegitimate unilateral sanctions' has been a test of strength for Russian compatriots abroad."	Hidden assumpt.: 1.0/ 0.0/ 2.5 Subjectivity: 1.4/ 1.0/ 2.3 Framing Russia: 0.0/ 0.0/ 0.0 Framing Ukraine: 0.0/ 0.0/ 0.0 Framing West: 0.0/ 0.0/ 0.0 Bias Pro-Russia: 1.2/ 1.0/ 1.0 Bias Pro-West: 0.8/ 1.0/ 1.0

a high degree of significance of our results, each article was annotated on a sentence-level by five crowdworkers. On average, they received about 3 to 4 cents per sentence annotation. Crowdworkers were presented with the whole article to have some context information when annotating sentences. As a consequence, they were paid per article depending on the article length.

4.2.1 Annotation Design. We used the following scales for the single crowdsourcing tasks:

- Hidden Assumptions: no; rather no; rather yes; yes; (0.0 to 3.0)
- Subjectivity: objective; rather objective; rather subjective; subjective; (0.0 to 3.0)
- Framing (for each government, i.e., Russian/Ukrainian/Western government(s)): negative; slightly negative; neutral; slightly positive; positive; (-2.0 to 2.0)
- Bias (for each tendency of bias, i.e., Pro-Russia/Pro-West): no; rather no; rather yes; yes; (0.0 to 3.0)

Overall, we spent $3,335 for the annotations. The subjectivity annotation had the lowest costs of $264, whereas the annotation of *bias* itself was most expensive with an amount of $1,322.

4.2.2 Quality Control. We used test questions (leftover news articles from Cremisini *et al.* [6]) and crowdworker experience to improve the annotation quality. We refrained from hard distinctions and allowed users to answer for certain test questions in multiple ways. All prepared answers were reviewed by four researchers.

4.3 Post-Processing

In the following, we outline how sentence and article labels were determined based on the crowdworkers' judgements.

4.3.1 Sentence Labels. Given that we have obtained judgements from five crowdworkers for the same sentence, there exists several ways to aggregate the answers (referred to as *calculation modes* in the following):

(1) The **majority vote** is the usual method to aggregate several judgements. The label with the most votes is ultimately taken.
(2) The **average vote** is widely used for crowdsourcing as well. It takes the average value of all answers as the final label.
(3) The **intensified vote** is a newly introduced method that takes the average of all non-neutral answers, given that at least two crowdworkers provide a non-neutral answer. Otherwise the majority vote is taken. In this way, amplitudes in the annotation should be emphasized, allowing for sensitivity to subtle annotations.

Table 2 shows two sentences from our data set with obtained labels. The total sentence label distribution regarding binary labels (i.e., *biased*/*non-biased* based on sentence score calculated by average vote) is shown in Table 5.

4.3.2 Article Labels. To obtain scores on the *article* level, we calculated for each article the relative ratios of sentences concerning each bias tendency. The same procedure was used for the bias dimensions subjectivity and hidden assumptions. For framing, each article received the proportions of negative sentences, neutral sentences, and positive sentences for each target of bias (government). We refrained from taking the simple ratio of all "framed" sentences, since this would result in positive and negative articles with the same amount of "framed" sentences to be labeled the same. As sentence labels are calculated by three different calculation modes, each article also received three different values.

4.4 Data Provisioning

After having all 2,057 sentences annotated with the four different labels (see Sec. 4.2.1) and aggregated by three different calculation modes (see Sec. 4.3.1), we obtained 43,197 sentence labels in total. In addition, aggregating the sentence labels for each article resulted in additional 1,350 article labels.

Our data set containing all labels is available online at **https://doi.org/10.5281/zenodo.3885351** and via our repository.[3] It is licensed under CC BY-NC 4.0 and can be reused for research purposes. The labels for each news article are provided in a separate file.

To ensure that our data set can be reused by other researchers and practitioners, we follow the *FAIR Guiding Principles for scientific data management and stewardship* [23]. These guidelines, which are applied widely nowadays, were designed to make resources **f**indable, **a**ccessible, **i**nteroperable, and **r**e-usable. In the context of our data set, these requirements are met as follows: (1) We provide a detailed description of the data set online.[4] (2) We provide the source-code for processing the data set (see Sec. 5) online for reproducibility and further usage.[5] (2) We added licensing information concerning the data set on our website, as well as to the manual in the data set. (3) We uploaded the data set on Zenodo and thereby made sure that the data set can be referenced permanently via a unique and resolvable DOI. Zenodo stores data at the CERN Data Center. Thus, we can rely on a long-term preservation of the data.

5 ANALYSIS OF THE DATA SET

For further investigation, we mapped all calculated labels to a binary score resulting in sentences and articles being classified as "biased" and "not biased." Table 3 indicates the number of articles

[3] See https://github.com/michaelfaerber/ukraine-news-bias.
[4] See https://doi.org/10.5281/zenodo.3885351.
[5] See https://github.com/michaelfaerber/ukraine-news-bias.

Table 3: Number of articles in which no single sentence was labeled as biased w.r.t. each bias dimension/tendency. For sentence-level information, see Table 5.

	Average	Majority	Intensified
Hidden Assumptions	78 / 90	88 / 90	72 / 90
Subjectivity	66 / 90	77 / 90	61 / 90
Framing: Russian Government	79 / 90	88 / 90	87 / 90
Framing: Ukrainian Government	76 / 90	90 / 90	86 / 90
Framing: Western Governments	78 / 90	89 / 90	89 / 90
Bias: Pro-Russia	69 / 90	89 / 90	89 / 90
Bias: Pro-West	72 / 90	90 / 90	90 / 90

not receiving a "biased" label with respect to the single bias dimensions and tendencies (grouped by calculation mode). A sentence is defined as "biased" if it receives a score greater than 1.0 on a 4-point scale from 0.0 (being neutral) to 3.0 (being biased) regarding subjectivity and hidden assumptions. Regarding bias itself, a sentence is "biased" if at least one of the two tendencies (i.e., pro-West, pro-Russia) receives a score greater than 1.0 on the same 4-point scale. For framing, a sentence is "biased" (or "framed") if at least one government receives a non-neutral score. Articles are regarded as "biased," if they contain at least one biased sentence. Considering the majority mode, 96% of all articles were not labeled as "biased" after averaging over all bias dimensions. In contrast, if we consider the intensified mode, 91% of all news articles become "not biased". This shows that using the intensified mode has some effect, but not as strong as one might expect. Thus, for the following correlation calculations, we selected article and sentence values calculated *by average mode*.

We calculated the *inter-annotator agreement* using Krippendorff's Alpha, due to its fit for tasks related to emotions or opinions. Agreement per bias dimension and tendency are shown in Table 4. *On sentence level*, the annotator agreement is low for each bias dimension. However, it is important to keep in mind that *experts* also disagree to a certain extent in sentence classification tasks [3]. Interestingly, the annotation of bias itself reached the smallest agreement among crowdworkers. This supports our decision to further investigate factors that influence judgements, such as the origin of crowdworkers (see Section 5.4). Apart from that, the general low agreement aligns with *not* choosing the majority vote as the only mechanism for sentence label calculation. *On the article level*, we calculated the average agreement score of each article. Regarding the annotations of framing (0.81), hidden assumptions (0.27), and bias itself (0.33), the average agreement is higher compared to the agreement score based on all sentences. Interestingly, all articles have either a very high (total agreement, i.e., alpha=1) or a low agreement score (defined as alpha < 0.1).

5.1 Correlations on Sentence Level

We chose Spearman's correlation coefficient over Pearson's correlation coefficient because the latter is more sensitive to outliers. Moreover, Pearson's correlation coefficient only describes a linear dependency between variables. We use the widely applied range rule of thumb to interpret correlation coefficients.

Table 4: Inter-annotator agreement using Krippendorff's α.

	Hidden Assumpt.	Subject-ivity	Framing	Bias
Neutral Leaning	0.19	-0.01	0.04	-0.02
Pro-Russia Leaning	0.12	0.07	0.08	-0.03
Pro-West Leaning	0.05	-0.03	0.00	-0.09
All Article Sentences	0.11	0.02	0.04	-0.05
Average of All Articles	0.27	0.01	0.81	0.33
Articles w/ Total Agreement	34.4%	11.1%	81.1%	42.2%
Articles w/ Low Agreement	65.6%	85.6%	15.6%	55.6%

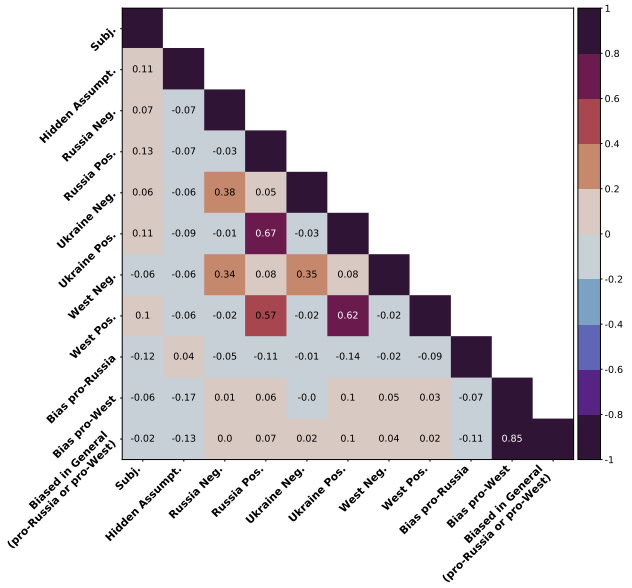

Figure 2: Correlations of bias dimensions on sentence level. A government and representation tendency (e.g., "Ukraine Neg.") is assigned to each framing dimension.

With respect to calculated sentence labels of the bias dimensions subjectivity and hidden assumptions, we can directly use the calculated labels of the post-processing step. The same applies to each bias tendency (i.e., pro-Russia/pro-West). For the dimension of framing, we created a positive and negative framing label indicating the intensity of their respective judgements in the sentence. In this way, we were able to analyze correlations between positive and negative representations independently.

5.1.1 Correlation of Bias Dimensions. Figure 2 shows the correlation coefficients for all bias dimensions and tendencies on the sentence level. We observe a weak correlation between all negative representations of governments. This means that the probability is higher than chance to find negative judgments toward one government if another government is already judged negatively in the same sentence. Interestingly, all positive representations have a moderate or high correlation.

5.1.2 Correlation between Bias Dimensions and Expert Labels. We also computed correlations between the sentence labels and the expert labels which classified the overall article leaning as being

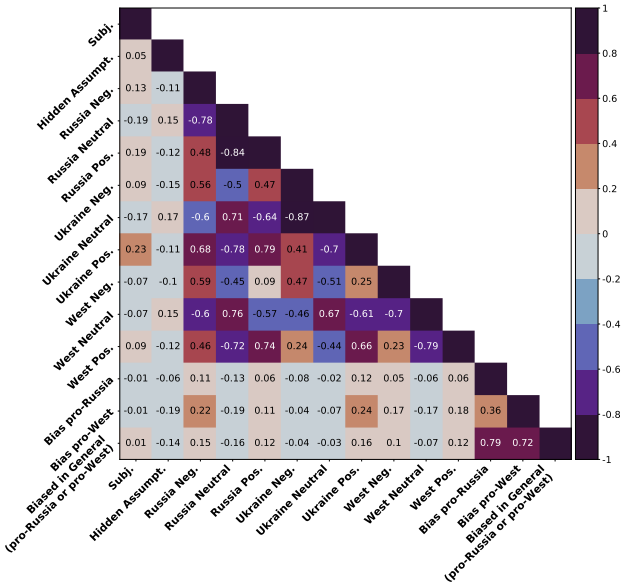

Figure 3: Correlations of bias dimensions on article level.

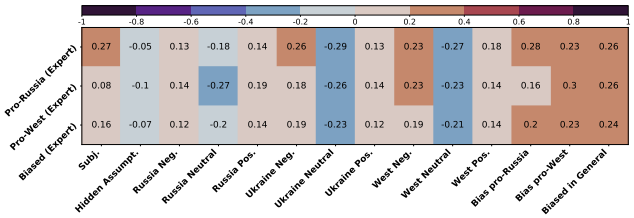

Figure 4: Correlation coefficients of bias dimensions and expert leaning on article level.

either pro-Russia, pro-West or neutral. Apart from pro-Russian and pro-West leaning articles, all non-neutral articles as defined by expert are grouped together as "biased" leaning articles. Similarly, crowdworkers' annotation of bias tendencies *pro-Russia* and *pro-West* is grouped together as "biased in general". We could not find any moderate or high correlation between any sentence labels and leaning groups. A weak correlation is found between the expert label "pro-Russia" and "biased in general". In contrast, there is also a very low correlation between the same expert label and "pro-West" on the crowdworkers' side. Thus, the leaning categorization of the expert seems not to be extractable qualitatively on a sentence level by bias dimensions/tendencies introduced via crowdsourcing.

5.2 Correlations on Article Level

5.2.1 Correlation of Bias Dimensions. Figure 3 shows the correlation matrix of obtained article labels. We did not find any correlation concerning SUBJECTIVITY and HIDDEN ASSUMPTIONS labels. However, we can observe a high correlation for the *neutral/positive* representations of all governments. Thus, if an article is comprised of many positive judgments of one government, the probability is high to find many positive judgements of another government. The same applies to negative judgements to a smaller extent because the negative representations are correlated with one another on a moderate level.

Additionally, several *opposite* representations are correlated with each other. For instance, the *negative* representation of the *Russian* government correlates on a high level with the *positive* representation of the *Ukrainian* government. This finding stands in contrast to the correlations calculated on sentence level. Moreover, it aligns with the general impression that the Russian and Ukrainian governments are antagonists in Ukraine Crisis.

For *bias itself*, we found a low correlation between both tendencies (see *Bias pro-Russia* and *Bias pro-West*). This suggests that crowd-workers had difficulties in annotating bias precisely with respect to the tendency in contrast to detecting an overall bias.

5.2.2 Correlation between Bias Dimensions and Expert Labeling. We examined the article leaning given by the expert annotation [6] and our inferred bias labels on article level (see Figure 4). Similarly to our analysis on the correlations between sentence and expert labels (see Sec. 5.1.2), we grouped pro-Russian and pro-West leaning articles (as defined by expert) together in a third leaning group in addition to each aforementioned leaning itself. Likewise, labels of both bias tendencies based on crowdworkers' annotations are also taken together as *biased in general*. In contrast to the results on the sentence level, we find a low correlation between pro-Russian/pro-West leaning articles and their corresponding crowdworker label. Controversially, pro-Russian articles also slightly correlate with *pro-West* labels given by crowdworkers.

With regard to the FRAMING dimension, pro-Russian articles show a higher probability to have a non-neutral view on the Ukrainian and Western governments. They also slightly correlate with the negative representation of the Ukrainian government. Similarly, pro-West leaning articles are more likely to contain a non-neutral view on the Ukrainian and Russian governments.

These findings all align again with the impression of (1) the Russian and Ukrainian government being antagonists and (2) the West being an additional opponent to Russian forces.

5.3 Distribution of Biased Sentences

Given all bias labels at the sentence level, we can analyze in which parts of the news articles certain bias labels (bias dimensions and bias itself) occur the most. Figure 5 shows the amount of biased sentences for each bias dimension and tendency regarding all possible (relative) positions within the article. We can observe the following:

Hidden Assumptions. HIDDEN ASSUMPTIONS are distributed relatively equally across the article parts. Compared to the other bias dimensions, we cannot observe significant outliers.

Subjectivity. Subjective sentences occur more often in the second half of an article.

Framing. The FRAMING dimension shows the most uneven distribution in all the news articles. 22 "framed" sentences are located in the interval of 60-80% of the content whereas 46 "framed" sentences are found in the last fifth part of the articles. We suspect it is because journalists tend to put evaluative statements on governments as targets of bias rather at the end of an article.

Bias. Interestingly, the direct bias dimension is the only dimension with the maximal amount occurring in the first fifth of an article. One likely reason for this phenomenon is the fact that the

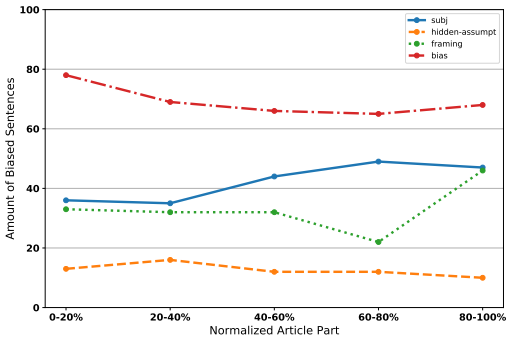

Figure 5: Amount of biased sentences w.r.t. their relative position within the articles.

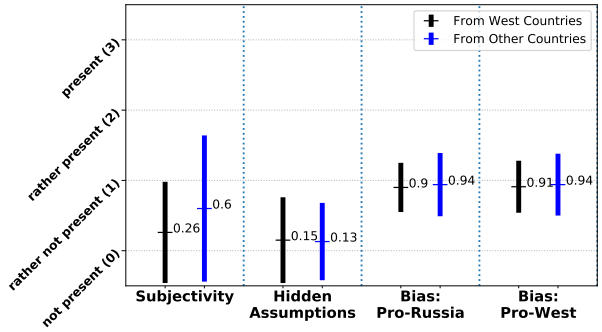

Figure 6: Average of crowdworkers' judgements with respect to their origins.

introducing summaries at the beginning of most news articles are particularly vulnerable to bias.

Overall, analyzing the occurrence of bias within the news articles is worth consideration in future work. To the best of our knowledge, this is the first analysis on where in articles different bias-related aspects can occur, and it was possible thanks to the fine-grained, sentence-level annotations we produced.

5.4 Influence of Crowdworkers' Origins

As laypersons usually consume news in their everyday lives and are surrounded continuously by media bias, the question arises whether certain groups of people have a different view on selected bias dimensions, including bias itself. Hence, we analyze the judgements of different crowdworker groups defined by their country of origin. In total, 570 crowdworkers participated in the data annotation. 220 crowdworkers came from Western countries, such as the USA, Spain, and the UK. From the remaining countries (excluding Post-Soviet countries), a total of 346 crowdworkers participated in the data annotation. Given this distribution and the Ukraine crisis as the news articles' topic, we compare those two crowdworker groups in the following.

5.4.1 Average Scores. The average scores of each dimension/tendency for the two crowdworker groups are shown in Figure 6 (except for framing due to average scores between 0.0 and 0.01 and similarly small standard deviations for both groups). We can observe the largest difference between the crowd-worker groups with respect

Table 5: Amount of Sentences recognized as biased w.r.t. bias dimension/tendency and crowdworkers' origin.

	All	From Western Countries	From Other Countries	Overlaps btw. West & Others
Hidden Assumptions	3.01%	4.02%	5.75%	0.65%
Subjectivity	10.36%	**4.07%**	**21.39%**	0.15%
Framing: General	8.02%	3.07%	6.96%	**1.98%**
Russian Gov. Pos.	2.67%	0.36%	2.71%	0.19%
Russian Gov. Neg.	2.24%	1.38%	1.64%	0.32%
Ukrainian Gov. Pos.	2.92%	0.24%	3.02%	0.19%
Ukrainian Gov. Neg.	2.48%	1.51%	1.54%	0.45%
Western Gov. Pos.	1.75%	0.42%	1.48%	0.00%
Western Gov. Neg.	1.46%	1.08%	0.61%	0.00%
Bias: General	16.53%	**6.05%**	**11.67%**	**0.44%**
Bias: Pro-Russia	9.09%	2.12%	8.90%	0.37%
Bias: Pro-West	12.06%	3.93%	7.34%	0.00%

to subjectivity. Crowd-workers from non-Western countries tend to find more subjective characteristics in all sentences. The average of their judgements regarding subjectivity is more than twice the average of all Western judgements. As for other dimensions we did not find a discrepancy so high between the crowdworker groups.

5.4.2 Amount of Biased Sentences. Table 5 shows the percentage of biased sentences for each bias dimension and crowdworker group. The overlap between both crowdworker groups is calculated as the ratio of sentences being labeled as biased by both groups to the amount of sentences being judged by both groups. Overall, the overlap between West and Others exceeds 1% only for *framing in general*, regardless of the fact that other dimensions received a greater proportion of biased sentences by both groups. The highest amount of biased sentences, according to Western crowdworkers, was detected by direct bias labeling with 6.05%. Even for this dimension, there is only an overlap of 0.44%. On top of that, crowdworkers from other than Western countries found more sentences being biased with regard to all dimensions, except for one out of the six variations of the *framing* dimension. For instance, the amount of subjective sentences given by non-Western crowdworkers is five times higher than the amount of subjective sentences according to Western crowdworkers. We hypothesize that Western people have already gotten more in touch with news on the Ukraine crisis, making it more difficult to keep an objective view. In conclusion, crowdworkers do show high differences in bias annotation depending on their origin – with regard to selection of biased sentences and sensitivity to bias.

6 IMPACT AND USE CASES

On the one hand, news articles remain the primary source to stay informed and to form opinions [8]. On the other hand, media bias has a strong impact on the individual and public perception of news topics leading to political changes [24]. Therefore, media bias issue must not be underestimated. Nearly all news consumers are affected by media bias [8]. Despite the fact that media bias analysis has a long tradition, analyzing media bias computationally with text mining methods and computational linguistics methods, as

well as developing approaches to detect news bias automatically, has picked up speed only in the recent years.

We believe that our data set will have a high impact in research and innovation. Due to the nature of media bias covering fields such as natural language processing, machine learning, psychology, and social sciences, we expect that our data set will be well received and reused in these diverse scientific disciplines. The fact that similar data sets on media bias have been published recently (see Section 2) indicates the need for data sets in this research area. Moreover, the high number of teams participating in the SemEval 2019 Task 4 [10] (322 registered & 42 participating teams) and the rising number of workshops and tracks on bias and fairness also support the claim that our data set and the communities using it will increase in the next few years.

We can think of several application scenarios for our data set:

News Bias Analysis. As demonstrated in Section 5, our data set can be used for an in-depth analysis of media bias in various regards. Particularly noteworthy is the fact that our data set contains, by far, the most bias-related labels on sentence level compared to existing data sets. Furthermore, the data set contains, to the best of our knowledge for the first time, several bias dimension labels per sentence.

Researchers can extend our data set with bias labels for news articles dealing with additional events or written in other languages. In this way, the backgrounds of authors and readers (e.g., gender, race, ethnicity, or language) can be studied.

News Bias Detection. Current news bias detection approaches (e.g., [10]) are lacking large evaluation and training data sets. With our data set, such approaches can be trained and evaluated on a considerably larger scale, making evaluations more trustworthy and findings more significant. Media bias detection systems can be integrated into *news recommender systems* and *news aggregation portals*. Also, users can be guided with respect to news bias via browser plug-ins (e.g., when reading online news [4]).

7 CONCLUSION

In this paper, we presented a new data set encompassing 43,197 sentence labels and 1,350 calculated article labels with respect to bias occurring in news articles. Based on a novel annotation schema which takes *hidden assumptions and premises*, *subjectivity*, *framing*, and *bias* itself into account. The data set facilitates an analysis of the perception of bias and related aspects. Our analysis of the data revealed several findings, which can be summarized as follows:

- *Average vote* outperforms majority vote as label calculation mode. Hence, we recommend it for similar sentence classification tasks.
- *Articles* received either very high or very low annotator agreement. Overall *sentence* agreement was low for each bias dimension indicating the difficulty of the annotation task.
- The bias dimension FRAMING seems to correlate on a low level with bias in news articles annotated by experts.
- Inferred *article* labels showed higher correlations to article labels of experts than *sentence* labels. Thus, bias detection systems might be more effective on article level rather than sentence level.
- The origin of crowdworkers affects their judgements. Non-Western persons found more bias either directly or in form of a bias dimension compared to Western people. This applies especially to their subjectivity annotation. Both crowdworker groups also perceived different sentences as biased.

In the future, we plan to use crowdsourcing for annotating news articles on a *word level* with respect to the bias dimensions and bias itself. Secondly, we will concentrate even more on the different backgrounds (e.g., gender, race, ethnicity) of authors and readers.

ACKNOWLEDGMENTS

This work was carried out with the support of the Baden-Württemberg Ministry of Science, Research and the Arts within the research project *digilog@bw*. We acknowledge support by the KIT-Publication Fund of the Karlsruhe Institute of Technology.

REFERENCES

[1] Alexandra Balahur et al. 2013. Sentiment Analysis in the News. *CoRR* abs/1309.6202 (2013). arXiv:1309.6202
[2] Eric Baumer, Elisha Elovic, Ying Qin, Francesca Polletta, and Geri Gay. 2015. Testing and Comparing Computational Approaches for Identifying the Language of Framing in Political News. In *Proc. of NAACL'15*. Denver, Colorado, 1472–1482.
[3] Kenneth Benoit et al. 2016. Crowd-sourced text analysis: Reproducible and agile production of political data. *Amer. Polit. Sci. Rev.* 110, 2 (2016), 278–295.
[4] Bjarte Botnevik, Eirik Sakariassen, and Vinay Setty. 2020. BRENDA: Browser Extension for Fake News Detection. In *Proc. of SIGIR'20*. 2117–2120.
[5] Ceren Budak, Sharad Goel, and Justin M. Rao. 2016. Fair and Balanced? Quantifying Media Bias through Crowdsourced Content Analysis. 80 (2016), 250–271.
[6] Andres Cremisini, Daniela Aguilar, and Mark A. Finlayson. 2019. A Challenging Dataset for Bias Detection: The Case of the Crisis in the Ukraine. In *Proc. of SBP-BRiMS'19*. 173–183.
[7] Robert M Entman. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of communication* 43, 4 (1993), 51–58.
[8] Felix Hamborg, Karsten Donnay, and Bela Gipp. 2019. Automated identification of media bias in news articles: an interdisciplinary literature review. *Int. J. on Digital Libraries* 20, 4 (2019), 391–415.
[9] Felix Hamborg, Anastasia Zhukova, and Bela Gipp. 2019. Automated Identification of Media Bias by Word Choice and Labeling in News Articles. In *Proc. of JCDL'19*. Champaign, IL, USA, 196–205.
[10] Johannes Kiesel et al. 2019. SemEval-2019 Task 4: Hyperpartisan News Detection. In *Proc. of SemEval@NAACL-HLT'19*. Minneapolis, MN, USA, 829–839.
[11] Sora Lim, Adam Jatowt, Michael Färber, and Masatoshi Yoshikawa. 2020. Annotating and Analyzing Biased Sentences in News Articles using Crowdsourcing. In *Proc. of LREC'20*. 1478–1484.
[12] Sora Lim, Adam Jatowt, and Masatoshi Yoshikawa. 2018. Understanding Characteristics of Biased Sentences in News Articles. In *Proc. of the CIKM2018 Workshops*.
[13] Fred Morstatter et al. 2018. Identifying framing bias in online news. *ACM Transactions on Social Computing* 1, 2 (2018), 1–18.
[14] Ndapandula Nakashole and Tom M. Mitchell. 2014. Language-Aware Truth Assessment of Fact Candidates. In *Proc. of ACL'14*. 1009–1019.
[15] Tatsuya Ogawa, Qiang Ma, and Masatoshi Yoshikawa. 2011. News Bias Analysis Based on Stakeholder Mining. E94-D, 3 (2011), 578–586.
[16] Souneil Park et al. 2011. NewsCube 2.0: An Exploratory Design of a Social News Website for Media Bias Mitigation. In *Proc. of SRS'11*.
[17] Thomas E. Patterson and Wolfgang Donsbagh. 1996. News decisions: Journalists as partisan actors. *Political Communication* 13, 4 (1996), 455–468.
[18] Gordon Pennycook and David G. Rand. 2018. Crowdsourcing Judgments of News Source Quality. (2018). https://doi.org/10.2139/ssrn.3118471
[19] Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2017. Automatic Detection of Fake News. *CoRR* abs/1708.07104 (2017).
[20] Martin Potthast et al. 2018. A Stylometric Inquiry into Hyperpartisan and Fake News. In *Proc. of ACL'18*. Melbourne, Australia, 231–240.
[21] Marta Recasens et al. 2013. Linguistic Models for Analyzing and Detecting Biased Language. In *Proc. of ACL'13*. 1650–1659.
[22] William Yang Wang. 2017. Liar, Liar Pants on Fire: A New Benchmark Dataset for Fake News Detection. In *Proc. of ACL'17*. 422–426.
[23] Mark D. Wilkinson et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3, 1 (2016), 1–9.
[24] John R. Zaller. 1992. *The Nature and Origins of Mass Opinion*. Cambridge University Press.
[25] Amy X. Zhang et al. 2018. A Structured Response to Misinformation: Defining and Annotating Credibility Indicators in News Articles. In *Proc. of WWW'18*. Lyon, France, 603–612.
[26] Jianwei Zhang et al. 2011. Sentiment Bias Detection in Support of News Credibility Judgment. In *Proc. of HICSS'11*. Koloa, Kauai, HI, USA, 1–10.