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.