Preprint from https://www.gipp.com/pub/

T. Spinde et al. "Towards A Reliable Ground-Truth For Biased Language Detection". In: *Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)*. Virtual Event, Sept. 2021. DOI: 10.1109/JCDL52503.2021.00053

Towards A Reliable Ground-Truth For Biased Language Detection

Timo Spinde
Dept. of Computer and Information Science
University of Konstanz
Constance, Germany
timo.spinde@uni-konstanz.de

David Krieger
Dept. of Computer and Information Science
University of Konstanz
Constance, Germany
jan-david.krieger@uni-konstanz.de

Manuel Plank
Dept. of Computer and Information Science
University of Konstanz
Constance, Germany
manuel.plank@uni-konstanz.de

Bela Gipp
School of Electrical, Information and Media Engineering
University of Wuppertal
Wuppertal, Germany
gipp@uni-wuppertal.de

***Abstract*—Reference texts such as encyclopedias and news articles can manifest biased language when objective reporting is substituted by subjective writing. Existing methods to detect bias mostly rely on annotated data to train machine learning models. However, low annotator agreement and comparability is a substantial drawback in available media bias corpora. To evaluate data collection options, we collect and compare labels obtained from two popular crowdsourcing platforms. Our results demonstrate the existing crowdsourcing approaches' lack of data quality, underlining the need for a trained expert framework to gather a more reliable dataset. By creating such a framework and gathering a first dataset, we are able to improve Krippendorff's α = 0.144 (crowdsourcing labels) to α = 0.419 (expert labels). We conclude that detailed annotator training increases data quality, improving the performance of existing bias detection systems. We will continue to extend our dataset in the future.**

***Index Terms*—Media Bias, News Slant, Dataset, Survey, Crowdsourcing**

I. Introduction and Related Work

The way journalists report on newsworthy events can influence consumers in their perception of political issues. Slanted coverage, also known as *media bias*, appears in different forms and on various linguistic levels [1]. The present project deals with the exploration of biased language on a word and sentence level.

Several studies have presented systems to detect slanted news reporting. Efforts include traditional machine learning classifiers relying on manual feature-engineering as in [2], [3], and neural-based methods [4]. To train and evaluate these algorithms, instances of text with a bias-inducing word choice or framing need to be labeled [5]. The need for training and validation data can be addressed by designing a crowdsourcing task as in [6]. The dataset created in this study via *Amazon Mechanical Turk* (MTurk) is the most exhaustive sample containing news bias labels on a fine-grained level to the best of our knowledge.[1] Yet, one of the approaches' main shortcomings is the resulting poor data quality concerning inter-rater reliability (IRR), which might negatively affect the performance of downstream classification tasks. Machine learning algorithms need rich training signals to learn an accurate language representation.

Crowdsourcing via MTurk has shown several drawbacks in past research: known problems are practice effects and the existence of discussion boards, resulting in a reduced naivety of the users. In contrast, those problems are not found to be existent to the same extent on other crowdsourcing platforms such as *Prolific* [7].[2] In further comparative annotation studies, MTurkers were less naïve and more familiar with the presented tasks than Prolific users. Beyond that, MTurkers showed a higher cheating rate than Prolific participants [8].

Our work aims to facilitate further research on the language conveying bias by elaborating on different ways to get fine-grained and qualitative annotations of biased language. As a first step towards compiling a reliable ground-truth for biased language detection, we compare IRR scores regarding media bias annotations on both MTurk and Prolific. Thereupon, we let trained experts label bias instances using detailed annotation instructions.

We summarize our hypotheses as follows: We assume that Prolific crowdsourcers show a higher agreement than MTurkers due to the presented drawbacks of the MTurk platform. Beyond that, we presume that expert training through detailed labeling instructions increases the annotation accordance.

II. Methodology and Results

A. Crowdsourcing

We first seek to compare user performances on MTurk and Prolific in the context of media bias. We rely on the news bias dataset provided by [6] as data ground-truth. It comprises 1,700 sentences with news bias annotations on word and sentence level extracted from 1000 articles. The dataset covers news platforms from the whole political spectrum. Furthermore, the survey includes a wide range of controversial topics with a balanced sociodemographic user characteristics distribution. The dataset being representative is crucial for the development of a generalizable bias detection tool.

We draw a representative sample of 100 sentences from the existing dataset [6]. The sample's characteristics are illustrated

This work was supported by the German Academic Exchange Service and the Hanns-Seidel-Foundation.

[1]https://www.mturk.com/

[2]https://www.prolific.com/

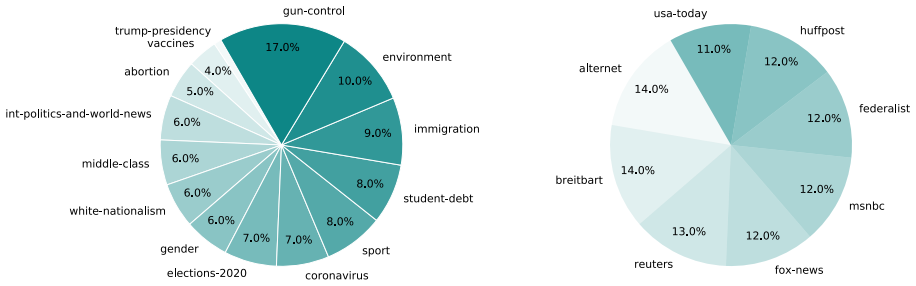

Fig. 1: Topic (left) and News Outlet (right) Distribution

in Fig. 1. To obtain a clear comparative picture we compare performance on MTurk and Prolific by asking Prolific users to re-annotate the existing MTurk labels. As in the original study, we ask the annotators to label the sentences in terms of bias on a sentence level and a word level. Our agreement metric of choice is Krippendorff's α [9]. The computational principle of this metric involves the ratio of the observed disagreement against the expected disagreement. We focus on the agreement on sentence level since including one linguistic level suffices for our crowdsourcing comparison. In future approaches, we also aim to analyze user agreement on a more fine-grained level. The agreement in the original study conducted on MTurk was $\alpha = .101$. Annotators on Prolific reach an $\alpha = .144$. The average number of clickworkers were 10.43 and 12.69 in the MTurk and Prolific study, respectively.

The Profilic users show a more profound agreement overlap in their bias ratings than the MTurkers. Yet, both agreement scores are by far not satisfying since [9] suggests a minimum α = .667 as the lowest conceivable limit. No available dataset comes near to that margin. These findings support the notion that identifying biased language is a complex task. We assume that crowdsourcers do not have sufficient knowledge regarding the linguistic theory and manifestations of the media bias concept. As a logical next step, we want to include bias ratings of experts to enhance data quality.

B. Expert Annotations

We hypothesize that the low quality of the data is the result of limited time. Crowdsourcers might mostly not be able to render accurate labels for this complex task. To mitigate scant domain knowledge problems, we follow corpus linguistic practice and develop detailed annotation instructions for coders. Annotation guidelines and data can be found at https://zenodo.org/record/4625151. Providing methodological steps for human-coders is essential but cannot be addressed sufficiently in a crowdsourcing setting due to time constraints. For the first sample presented in this poster, we employ two annotators working in the context of media bias. Preliminary results on 1,700 sentences are encouraging, yielding an α of .419, which surpasses available datasets by a large margin. Exceeding the annotator pool to 12 annotators and experimenting with other users is a work-in-progress.

III. Conclusions and Future Work

This poster proposes a work-in-progress approach to compile a ground-truth dataset suitable for media bias detection. So far, we implemented a crowdsourcing comparison of user performances on the media bias detection task. We let users annotate an exemplary sentence corpus both on MTurk and Prolific. Prolific participants outperformed MTurkers with a Krippendorff's α = .144 vs. α = .101. We conclude that these low agreement scores are due to the crowdsourcers' insufficient understanding of the media bias concept.

Furthermore, including trained experts improved the data quality by increasing the annotators' agreement to an α = .419. As a next step, we plan to build a diverse team of annotators to improve the current dataset quantitatively and qualitatively. We expect that future computational detection approaches will benefit substantially from this development.

References

[1] Timo Spinde, Kanishka Sinha, Norman Meuschke, and Bela Gipp. Tassy - a text annotation survey system. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, Sep. 2021.

[2] Timo Spinde, Lada Rudnitckaia, Jelena Mitrović, Felix Hamborg, Michael Granitzer, Bela Gipp, and Karsten Donnay. Automated identification of bias inducing words in news articles using linguistic and context-oriented features. *Information Processing & Management*, 58(3):102505, 2021.

[3] Timo Spinde, Felix Hamborg, and Bela Gipp. Media bias in german news articles : A combined approach. In *Proceedings of the 8th International Workshop on News Recommendation and Analytics (INRA 2020)*, Virtual event, 2020.

[4] Christoph Hube and Besnik Fetahu. Neural based statement classification for biased language. In *Proceedings of the twelfth ACM international conference on web search and data mining*, pages 195–203, 2019.

[5] Timo Spinde, Christina Kreuter, Wolfgang Gaissmaier, Felix Hamborg, Bela Gipp, and Helge Giese. Do you think it's biased? how to ask for the perception of media bias. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, Sep. 2021.

[6] Timo Spinde, Lada Rudnitckaia, Sinha Kanishka, Felix Hamborg, Bela, Gipp, and Karsten Donnay. Mbic – a media bias annotation dataset including annotator characteristics. In *Proceedings of the iConference 2021*, Beijing, China (Virtual Event), 2021.

[7] Stefan Palan and Christian Schitter. Prolific. ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27, 2018.

[8] Eyal Peer, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. Beyond the turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, 70:153–163, 2017.

[9] Klaus Krippendorff. Content analysis: An introduction to its methodology 2nd thousand oaks, 2004.