Preprint from: https://media-bias-research.org/publications/

J. Krieger et al. "A Domain-adaptive Pre-training Appraoch for Language Bias Detection in News". In: *Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)*. Cologne,Germany, June 2022. DOI: 10.1145/3529372.3530932

A Domain-adaptive Pre-training Approach for Language Bias Detection in News

Jan-David Krieger*
jan-david.krieger@uni-konstanz.de
University of Konstanz
Konstanz, Germany

Timo Spinde*
timo.spinde@uni-wuppertal.de
University of Wuppertal
Wuppertal, Germany

Terry Ruas
ruas@uni-wuppertal.de
University of Wuppertal
Wuppertal, Germany

Juhi Kulshrestha
juhi.kulshrestha@uni-konstanz.de
University of Konstanz
Konstanz, Germany

Bela Gipp
gipp@cs.uni-goettingen.de
University Göttingen
Göttingen, Germany

ABSTRACT

Media bias is a multi-faceted construct influencing individual behavior and collective decision-making. Slanted news reporting is the result of one-sided and polarized writing which can occur in various forms. In this work, we focus on an important form of media bias, i.e. *bias by word choice*. Detecting biased word choices is a challenging task due to its linguistic complexity and the lack of representative gold-standard corpora. We present DA-RoBERTa, a new state-of-the-art transformer-based model adapted to the media bias domain which identifies sentence-level bias with an F1 score of 0.814. In addition, we also train, DA-BERT and DA-BART, two more transformer models adapted to the bias domain. Our proposed domain-adapted models outperform prior bias detection approaches on the same data.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**; • **Information systems** → **Clustering and classification**.

KEYWORDS

Media bias, news slant, neural classification, text analysis, domain adaptive

ACM Reference Format:
Jan-David Krieger, Timo Spinde, Terry Ruas, Juhi Kulshrestha, and Bela Gipp. 2022. A Domain-adaptive Pre-training Approach for Language Bias Detection in News. In *The ACM/IEEE Joint Conference on Digital Libraries in 2022 (JCDL '22), June 20–24, 2022, Cologne, Germany*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3529372.3530932

1 INTRODUCTION

Over the last few years, online news has increasingly replaced traditional printed news formats [7, 12, 17, 32]. Online news environment provides information from diverse sources with varying perspectives, thereby, allowing people to decide which source they want to consume [2]. Unfortunately, the diversity of online news sources also opens the door for slanted and non-neutral news coverage [36]. Biased news coverage - referred to as *media bias* in the literature [35, 42, 43] - occurs once subjective reporting on a specific event replaces objective coverage. Media bias manifests in various forms such as bias by word choice [45] or bias by omission [25] of information. For more examples of media bias, we refer to [45].

Detecting and potentially reducing media bias in the news is societally relevant on multiple accounts. For policy regulators and related organizations, automated bias detection can help keep a tab on the bias in different outlets in the news ecosystem. For news consumers, it can help the development of tools for mitigating any adverse effect of the media bias on them. For journalists, automatic bias identification can improve their writing through more objective reporting [45]. Ideally, in the future, journalistic writing tools would mitigate biases in news reporting by accurately prompting reporters when their news coverage exhibits linguistic bias.

Detecting media bias is a complex task due to its subtle nature and the lack of a clear and unique linguistic definition. Therefore, developing accurate quantitative detection approaches is known to be a challenging task in media bias research [11, 22, 39, 45]. One of the main problems in media bias research is the lack of exhaustive gold-standard bias datasets for pre-training large-scale language models (e.g., BERT [8]). Prior transformer-based approaches tackle the resource limit by incorporating bias-related datasets into pre-training techniques such as *Distant Supervision Learning* [42] and *Multi-task Learning* [41] yielding performance improvements in only some experimental setups. Respective studies either rely on noisy and marginally bias-related training data [42], or do not fully exploit highly bias-related data by incorporating only sub-samples of bias corpora into pre-training [41].

We propose an effective *domain-adaptive pre-training* approach that relies on a highly relevant bias-related encyclopedia data set. Similar approaches have been shown to yield substantial performance boosts for similar tasks within the news, biomedical, and scientific domains [4, 14, 15, 20, 47, 49]. To the best of our knowledge, domain-adaptive pre-training has not yet been explored in the media bias domain.

Our primary research objective is to assess the effects of domain-adaptive pre-training on the media bias detection performance of several large-scale language models. Our key contribution is

*Both authors contributed equally to this research.

JCDL '22, June 20–24, 2022, Cologne, Germany

ACM ISBN 978-1-4503-9345-4/22/06.
https://doi.org/10.1145/3529372.3530932

to leverage transformer-based models with an understanding of biased language. We perform an intermediate pre-training procedure with *BERT* [8], *RoBERTa* [24], *BART* [21], and *T5* [30] on the *Wiki Neutrality Corpus* (WNC) [29], which contains 180k sentence pairs from *Wikipedia* labeled as biased/neutral [29] and fine-tune the architecture on the state-of-the art media bias data set *BABE* [42]. We publish our domain-adapted models, i.e. *DA-RoBERTa* (DA = domain-adaptive), *DA-BERT*, *DA-BART*, and *DA-T5*, as well as training data and all material on https://github.com/Media-Bias-Group/A-Domain-adaptive-Pre-training-Approach-for-Language-BiasDetection-in-News. DA-RoBERTa achieves a new state-of-the-art performance on BABE (F1 = 0.814), while DA-BERT, DA-BART, and DA-T5 also outperform the baselines and distantly supervised models from prior work [42].

2 RELATED WORK

While media bias occurs in various forms (e.g, bias by omission, editorial bias) [45], our work focuses on bias by word choice induced by choosing different words to refer to the same concept [45]. A detailed introduction on different media bias forms can be found in Recasens and Jurafsky [31].

Several studies tackle the challenge of identifying biased language automatically. Early approaches used hand-crafted linguistic features to detect slanted news coverage on word- [31, 45] and sentence-level [16] based on traditional machine learning techniques. Since these approaches have been shown poor performance in bias detection, we do not experiment with manually generated bias-inducing features. Instead, we only include feature-based results from Spinde et al. [42] as a baseline in our experiments. A detailed introduction of feature-based bias detection studies can be found in Spinde et al. [42] .

In the rest of the section, we first summarize drawbacks of existing media bias corpora and justify why we focus on a single state-of-the-art bias corpus for evaluative purposes. Next, we discuss relevant transformer-based bias detection approaches and domain-adaptive pre-training studies.

2.1 Drawbacks of existing bias corpora

Several approaches tackle the challenge of creating representative media bias data sets [3, 22, 23, 42, 44, 46]. However, most corpora exhibit substantial drawbacks such as low inter-annotator agreement [3, 11, 22, 22, 44], low number of covered topics [23], or they focus on other concepts such as framing rather than on bias [3].

To the best of our knowledge, the most exhaustive media bias data set - BABE (Bias Annotations By Experts) Spinde et al. [42], contains 3700 sentences covering a wide range of topics and news articles from various news outlets. Five media bias experts labeled sentences in terms of bias on sentence- and word-level, among others. The resulting inter-annotator agreement on sentence-level is 0.39 measured by Krippendorff's α [19], which is much higher compared to other corpora. Therefore, we solely rely on BABE for our experimental evaluations, since no other bias data set exhibits similar data quality and representativeness. We plan to conduct more experiments applying our domain-adaptive approach in future datasets, assuming they will incorporate the aspects already present in BABE.

2.2 Transformer-based detection approaches

The linguistic subtlety of slanted news coverage is known to be a great challenge for automated classification methods [42]. Recent media bias studies have progressed from manually generated linguistic features [37, 38] to state-of-the-art NLP models yielding internal word representations by unsupervised or supervised training on massive text corpora. The Transformer architecture [48] has shown superior performance in several downstream tasks, such as, text classification [26–28], plagiarism detection [50, 51], word sense disambiguation [52] and fake news detection on the health domain [49]. However, the use of neural language models, such as BERT [8] and RoBERTa [24] in the media bias domain is still incipient [41, 42]. In this work, we contribute to mitigate this problem by applying the aforementioned language models via a domain-adaptive approach [14, 15, 47].

Spinde et al. [42] pre-train transformer-based models such as BERT [8], RoBERTa [24], and DistilBERT [34] using Distant Supervision Learning on news headlines from articles with different political leanings and fine-tune it on BABE [42]. Their best-performing models classify biased/non-biased sentences extracted from BABE with F1 scores of 0.804 (BERT) and 0.799 (RoBERTa). The authors also incorporate a feature-based classifier and show that transformer models substantially outperform the feature-based approach. As transformer-based models have been shown to clearly outperform feature-based ones, we exclude the latter from our experiments.

Spinde et al. [41] train DistilBERT [34] on combinations of bias-related datasets using a Multi-task Learning (MTL) [6, 54] approach. Their best-performing MTL model achieves 0.776 F1 score on a subset of BABE. However, the MTL model is outperformed by a baseline model (F1 = 0.782) trained on a subset of the datasets (WNC) used. Spinde et al. [41] suggest that improvements can be attributed to the WNC dataset being strongly bias-related, hence equipping the model with bias-specific knowledge.

While Spinde et al. [42] do not fully exploit bias-related data sets in their pre-training approach, Spinde et al. [41] implement a complex MTL architecture reducing the WNC's pre-training effect on the bias classification task. In our work, we use a similar learning task as Spinde et al. [42] and exploit the WNC's bias-relatedness by extending the pre-training of several transformer models on the whole WNC instead of its subset.

2.3 Domain-adaptive pre-training approaches

Our training setup can be considered a form of domain-adaptive pre-training [4, 15, 20] in which a language model is equipped with domain-specific knowledge. Several studies experiment with domain-adaptive learning approaches in different domains (e.g., BioBERT [20], SciBERT [4]), but none of them deals with media bias detection [4, 14, 15, 20, 47].

Sun et al. [47] explore different techniques for domain-adaptive pre-training of BERT for text classification tasks such as sentiment classification, question classification, and topic classification. BERT is additionally pre-trained on data from various domains leading to performance boosts on many tasks if the training data are related to the target task's domain. After training BERT on several sentiment classification datasets, Sun et al. [47] reduced the error

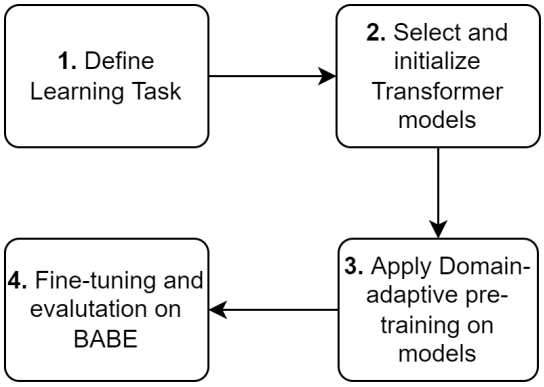

Figure 1: Pre-training and fine-tuning workflow.

rate on the Yelp sentiment data set to 1.87% (compared to 2.28% from BERT baseline initialized with *bert-base-uncased* weights). The results from Sun et al. [47] are supported by Gururangan et al. [14] investigating domain-adaptive pre-training of RoBERTa in four different target domains (i.e., biomedical, computer science publications, news, and reviews) and eight subsequent classification tasks. When pre-training RoBERTa on large amounts of news text, the model's F1-score on a hyperpartisan classification dataset [18] improves from F1 = 0.886 (*roberta-base* weights) to F1 = 0.882. Training the model on a domain outside the domain of interest (*irrelevant domain-adaptive pre-training*) drastically decreases performance to F1 = 0.764.

Our domain-adaptive pre-training approach is performed on the WNC corpus and based on implementations from Sun et al. [47] and Gururangan et al. [14]. Due to drastic performance increases through irrelevant domain-adaptive pre-training in previous research [14], we do not implement respective experiments. We detail our proposed training process and experiments in Sections 3 and 3.4. Since most existing approaches focus on sentence-level bias detection, we follow the standard practice and develop a sentence-level classification model. Compared to cutting-edge but convoluted studies in media bias detection [40, 42], we perform a more focused and direct training setup on a large amount of highly bias-related data and expect substantial performance improvements.

3 METHODOLOGY

We use neural-based language models, pre-train them on the bias domain (WNC), and perform evaluations on the media bias classification task using BABE as Figure 1 shows. We expect that domain-adaptive pre-training improves word representations by adapting them to the data distributions of biased and non-biased news content. Based on BABE, we define a learning task that is later optimized (Section 3.1). Then, we select suitable transformer models and initialize them with pre-trained weights (section 3.2). We adapt the models to the media bias domain by training them on the WNC (section 3.3). Finally, all models are fine-tuned and evaluated on BABE.

3.1 Learning task

The language models are optimized via intermediate training. We have a corpus X containing sentences $x_i \in X$ with $i = 1, ..., N$ and binary bias labels (*Biased* vs. *Non-biased*) encoded as 1 and 0, respectively. The task is to assign the correct label $y_i \in \{0, 1\}$ to x_i. The training objective is to minimize a binary cross-entropy loss

$$\mathcal{L} := -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=\{0,1\}} f_k(x_i) \cdot log(\hat{f}_k(x_i)). \quad (1)$$

where $f_k(x_i)$ refers to the true binary label and $\hat{f}_k(x_i)$ indicates the model's predicted score for a sentence.

3.2 Transformer-based models

We choose BERT and RoBERTa for our domain-adaptive pre-training as they represent the best-performing models in Spinde et al. [42]. Doing so, we also achieve maximum comparability to previous state-of-the-art bias classifiers. Additionally, we incorporate BART and T5, since encoder-decoder architectures have demonstrated a clear improvement in comparison to BERT in several NLP tasks (e.g., GLUE [53]). We choose the corresponding models to investigate how the combination of autoencoder and autoregressive components (BART), and advanced MTL architectures (T5) perform on our media bias detection task.

BERT learns bidirectional word representations on unlabeled text optimizing an unsupervised learning task based on *Masked Language Modeling* and *Next Sentence Prediction*. In contrast to BERT, RoBERTa drops the Next Sentence Prediction task and differs slightly in terms of pre-training data. BART uses text manipulations by noising and learns representations by reconstructing the original text sequence. T5 uses an MTL architecture pre-trained on various supervised and unsupervised tasks by converting all training objectives into text-to-text tasks. All models are adapted to the media bias domain (Section 3.3) and evaluated on the sentence-level media bias classification task (Section 3.4).

3.3 Domain-adaptive pre-training

Adapting a pre-trained language model to a specific domain becomes essential when the target domain differs strongly from the pre-training ground truth [4, 15, 20]. Due to tendentious and dubious vocabulary in slanted news, media bias is different from most of the domains BERT-like models are pre-trained on. For example, BERT is trained on English Wikipedia and the BooksCorpus [55] while RoBERTa additionally incorporates commonsense reasoning data, news data, and web text data. To the best of our knowledge, a specific BERT-like model trained on biased language in news does not exist to date. BERT models pre-trained on fake news [5] and political orientation classification [13] do exist. However, the concepts of fake news and political orientation differ substantially from the media bias domain.

Our domain-adaptive pre-training uses the WNC to optimize our learning task defined in Section 3.1. The 180k sentence pairs contained in the corpus are manually selected from Wikipedia articles as going against the platform's *Neutral Point of View* (NPOV) standard[1]. The pairs contain an original biased sentence and its

[1]https://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view

manually derived neutral counterpart. Bias forms included in the corpus refer to *epistemological bias*, *framing bias*, and *demographic bias*. Recasens and Jurafsky [31] define framing bias as choosing subjective words to embed a particular point of view in the text whereas epistemological bias is described as a modification of a statement's plausibility. Pryzant et al. [29] introduce demographic bias as text containing predispositions towards a certain gender, race, or other demographic category. For a detailed description on sentence selection criteria and the revision process, see Pryzant et al. [29].

Our approach is inspired by Sun et al. [47] and Gururangan et al. [14], which conclude domain-adaptive pre-training is most efficient once pre-training data for the domain adaption is related to the target domain and task. Since WNC (pre-training) and BABE (fine-tuning) have similar bias forms, and are both composed of manually labeled sentences (biased and non-biased), we expect the proposed pre-training task to improve our fine-tuning results.

3.4 Experiments

3.4.1 Pre-training. We initialize RoBERTa, BERT, BART, and T5 with pre-trained weights provided by the HuggingFace API[2], and stack a dropout layer (Dropout = 0.2) and randomly initialized linear transformation layer (768,2) on top of the model. All models are used in their base form.

For the domain-adaptive pre-training, sentences are batched together with 32 sentences per batch. For model optimization, we use the AdamW optimizer[3] with a learning rate of $1e^{-5}$, and model performance is evaluated on binary cross-entropy loss. Model convergence can be observed after one epoch and a runtime of $\approx$ 5 hours on a Tesla P100-PCIE GPU with 16GB RAM.

3.4.2 Fine-tuning. We fine-tune and evaluate the model on BABE Spinde et al. [42] with a batch size = 32. We again use the AdamW optimizer (learning rate = $1e^{-5}$), and model convergence based on cross-entropy loss can be observed after 3-4 epochs. Due to the small data size of 3700 sentences, we report the model's F1 score in the binary bias labeling task averaged by 5-fold cross-validation. Fine-tuning is performed on a Tesla K80 GPU (12GB RAM) in $\approx$ 15 minutes.

3.4.3 Baseline. For every domain-adaptive language model, we compare its sentence classification performance on BABE to the same architecture merely fine-tuned on BABE (without domain-adaptive pre-training as an intermediate training step). Thereby, we can assess the effect of our training approach. Since Spinde et al. [42] achieve state-of-the-art results on BABE with Distant Supervision Learning [42], we additionally compare our F1 scores to their scores achieved by training BERT and RoBERTa on news headlines distantly labeled as biased and non-biased. We provide statistical significance tests for our domain-adapted models vs. fine-tuned-only models.

3.4.4 Test for Statistical Significance. In their review on existing NLP studies, Dror et al. [10] report that most approaches lack statistical tests inspecting the significance of experimental results. The authors recommend various parametric and non-parametric test to compare performances of Machine Learning models.

For our approach, we select the *McNemarś test* which is a non-parametric test to compare the performance of two algorithms on a target task. Since we do not have information on the distribution of our target metric (F1 score), a non-parametric approach is a suitable option to test for significance. The test is based on a $2x2$ contingency table showing the models' predictions on n instances of a target taskś test set. Under the null hypothesis H_0, the test assumes that both algorithms output the correct/incorrect label for the same proportion of instances from the test set. Accordingly, the alternative hypothesis H_1 states that both algorithms differ significantly in terms of their agreement on items from the test set. The test statistic follows a χ^2 distribution and is suitable for NLP tasks such as binary text classification [9, 10]. For a more detailed introduction on statistical significance tests for NLP use cases, see Dror et al. [10].

4 RESULTS

Table 1 shows the F1 scores (averaged over 5-fold CV split) of our transformer-based experiments on the binary sentence classification task. All domain-adapted models (third block) outperform the baselines models (first block) and the distantly supervised models (middle block) trained by Spinde et al. [42].

The best-performing model that achieves a new state-of-the-art on BABE is DA-RoBERTa (F1 = 0.814), surpasses the baselines and its Distant Supervision variant by 1.5 %. DA-BERT, DA-BART, and DA-T5 achieve a lower F1-score of 0.809, 0.809, and 0.798, yet outperform BERT, BART, and T5 by 2%, 0.8%, and 1.2%, respectively. However, DA-BERT increases sentence classification performance by only 0.5% compared to BERT trained via Distant Supervision [42]. To the best of our knowledge, a distantly supervised variant for BART and T5 is not available.

Table 2 shows results of the McNemar statistical significance tests comparing our domain-adapted models with respective baselines. On a significance level of $\alpha = 0.05$, we can observe significant F1-score improvements for BERT vs. its domain-adapted variant ($\chi^2 = 5.65, p = 0.031$) as well as for RoBERTa vs. DA-RoBERTa ($\chi^2 = 3.844, p = 0.049$) and T5 vs. DA-T5 ($\chi^2 = 4.86, p = 0.027$). Adapting BART to the bias domain seems not to significantly improve the sentence classification performance ($\chi^2 = 3.629, p = 0.057$).

5 DISCUSSION

With DA-RoBERTa, we provide a new state-of-the-art classifier for the detection of biased language in the news articles on sentence-level. Furthermore, we show that all domain-adapted models outperform their baselines and distantly supervised models published by Spinde et al. [42]. Our results can be considered a contribution towards a sufficiently accurate bias detection tool. However, some significance tests comparing the performance of domain-adapted models vs. distantly supervised models are missing due to limited resources.

As indicated in Section 2.2, Spinde et al. [41] pre-train DistilBERT on a subset of the WNC and observe performance boosts of 3.6% on bias sentence classification with a subset of BABE compared to their baseline without intermediate pre-training. Although our

[2]https://huggingface.co/
[3]https://huggingface.co/docs/transformers/main_classes/optimizer_schedules

Table 1: Stratified 5 fold cross-validation results.

Model	Macro F1 (error)
BERT	0.789 (0.011)
RoBERTa	0.799 (0.011)
BART	0.801 (0.009)
T5	0.786 (0.008)
BERT-distant [42]	0.804 (0.014)
RoBERTa-distant [42]	0.799 (0.017)
DA-BERT	0.809 (0.010)
DA-RoBERTa	**0.814** (0.004)
DA-BART	0.809 (0.009)
DA-T5	0.798 (0.009)

Note: Standard errors across folds in parentheses.
The first block shows results of baseline approaches without intermediate pre-training. The second block shows results from [42] based on Distant Supervision Learning (BART and T5 are not incorporated in their study). Results from our domain-adaptive approach are shown in the third block.
The best result is printed in **bold**.

Table 2: Results of the McNemar test for statistical significance between baseline (without domain-adaptive pretraining) and domain-adapted models.

Models	McNemar test statistic	
	χ^2	p
BERT vs. DA-BERT	5.65	0.031*
RoBERTa vs. DA-RoBERTa	3.84	0.049*
BART vs. DA-BART	3.63	0.057
T5 vs. DA-T5	4.86	0.027*

Note: $*p < .05$

domain-adapted models incorporate the complete WNC into pretraining, we observe minor performance increases when compared to those obtained by DistilBERT. We believe that smaller-scaled and distilled models such as DistilBERT benefit more from additional pre-training than larger models relying on a different training objective such as BERT, RoBERTa, BART, and T5.

In the future, it will be interesting to verify how even more robust and general NLP models benefit from intermediate pre-training. Possibly, state-of-the art NLP models such as the recently published *ExT5* [1], incorporating extensive Multi-task Learning on 107 tasks from different domains, further decreases domain-adaptive learning effects. Furthermore, we expect that bias corpora such as BABE will continue to be proposed. From a resource consumption perspective, fine-tuning robust language models such as ExT5 on more representative bias corpora might be sufficient to achieve state-of-the-art performances in bias detection.

Considering our bias detection task, we want to point out that our models are merely trained to identify slanted news coverage on sentence-level. Since media bias is a linguistically complex construct [45], we need robust and more general classifiers for different linguistic bias perspectives such as word-, paragraph-, and article-level. As Recasens and Jurafsky [31] show, word-level detection of slanted news coverage is challenging for both humans and machines. Computer Science approaches dealing with bias on word-level might depend on collaborations with researchers from the Social Sciences to develop a large number of linguistically fine-grained gold-standard data for efficient model training. Furthermore, we need systems detecting various sub-forms of bias such as framing bias and epistemological bias accurately. MTL approaches trained on different bias categories might be a a promising direction for future models.

Future research should incorporate a broader range of evaluation tasks to assess model performance. Spinde et al. [42] argue that standard metrics such as F1 score are not sufficient to evaluate language models on the complex bias detection task. The authors suggest developing more advanced evaluation metrics such as decomposing the bias detection task into several subtasks to assess a model's detection power properly. Ribeiro et al. [33] introduce CheckList, a tool structuring the target task into several sub-tasks. Respective evaluation approaches could help assess a model's bias identification performance on different forms of bias.

6 CONCLUSION

This work proposes DA-RoBERTa, a new state-of-the-art language model for sentence-level detection of biased language in the news. We equip several transformer architectures (i.e., BERT, RoBERTa, BART, and T5) with an understanding of biased language, showing that domain-adaptive pre-training significantly improves the classifier's bias detection performance compared to baseline models without intermediate pre-training. Limitations of our approach are the exclusively pre-training focus on sentence-level classification and the restricted evaluation incorporating a single data set/task due to the lack of existing representative bias corpora. We hope that further high-quality bias corpora are published in the future to improve the generalizability of results and enable a more fine-grained and large-scale evaluation of models in the domain. Considering continuous developments in the NLP field, future studies should also address whether upcoming more robust language models still require intermediate pre-training on the media bias domain.

ACKNOWLEDGMENTS

The Hanns-Seidel-Foundation, Germany, supported this work, as did the DAAD (German Academic Exchange Service).

REFERENCES

[1] Vamsi Aribandi, Yi Tay, Tal Schuster, Jinfeng Rao, Huaixiu Steven Zheng, Sanket Vaibhav Mehta, Honglei Zhuang, Vinh Q. Tran, Dara Bahri, Jianmo Ni, Jai Gupta, Kai Hui, Sebastian Ruder, and Donald Metzler. 2021. ExT5: Towards Extreme Multi-Task Scaling for Transfer Learning. arXiv:2111.10952 [cs.CL]
[2] Mahmoudreza Babaei, Juhi Kulshrestha, Abhijnan Chakraborty, Fabrício Benevenuto, Krishna P. Gummadi, and Adrian Weller. 2018. Purple feed: Identifying high consensus news posts on social media. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. ACI, New York, 10–16.
[3] 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. 1472–1482. https://doi.org/10.3115/v1/N15-1171
[4] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of the 2019 Conference on Empirical*

Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics, Hong Kong, China, 3615–3620. https://doi.org/10.18653/v1/D19-1371

[5] Ciara Blackledge and Amir Atapour-Abarghouei. 2021. Transforming Fake News: Robust Generalisable News Classification Using Transformers. arXiv:2109.09796 [cs.CL]

[6] Shijie Chen, Yu Zhang, and Qiang Yang. 2021. Multi-Task Learning in Natural Language Processing: An Overview. *arXiv:2109.09138 [cs]* (Sept. 2021). http://arxiv.org/abs/2109.09138 arXiv: 2109.09138.

[7] Alexander Dallmann, Florian Lemmerich, Daniel Zoller, and Andreas Hotho. 2015. Media Bias in German Online Newspapers. In *Proceedings of the 26th ACM Conference on Hypertext amp; Social Media* (Guzelyurt, Northern Cyprus) *(HT '15)*. Association for Computing Machinery, New York, NY, USA, 133–137. https://doi.org/10.1145/2700171.2791057

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[9] Thomas G Dietterich. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation* 10, 7 (1998), 1895–1923. https://doi.org/10.1162/089976698300017197

[10] Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The Hitchhiker's Guide to Testing Statistical Significance in Natural Language Processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Melbourne, Australia, 1383–1392. https://doi.org/10.18653/v1/P18-1128

[11] Soumen Ganguly, Juhi Kulshrestha, Jisun An, and Haewoon Kwak. 2020. Empirical Evaluation of Three Common Assumptions in Building Political Media Bias Datasets. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 939–943.

[12] Marcel Garz and Gregory J. Martin. 2021. Media Influence on Vote Choices: Unemployment News and Incumbents' Electoral Prospects. *American Journal of Political Science* 65, 2 (2021), 278–293. https://doi.org/10.1111/ajps.12539

[13] Shloak Gupta, S Bolden, Jay Kachhadia, A Korsunska, and J Stromer-Galley. 2020. PoliBERT: Classifying political social media messages with BERT. In *Social, Cultural and Behavioral Modeling (SBP-BRIMS 2020) conference. Washington, DC.*

[14] Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. *CoRR* abs/2004.10964 (2020). arXiv:2004.10964 https://arxiv.org/abs/2004.10964

[15] Xiaochuang Han and Jacob Eisenstein. 2019. Unsupervised Domain Adaptation of Contextualized Embeddings for Sequence Labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 4238–4248. https://doi.org/10.18653/v1/D19-1433

[16] Christoph Hube and Besnik Fetahu. 2018. Detecting biased statements in wikipedia. In *Companion proceedings of the the web conference 2018*. 1779–1786. https://doi.org/10.1145/3184558.3191640.

[17] Barbara Kaye and Thomas Johnson. 2016. Across the Great Divide: How Partisanship and Perceptions of Media Bias Influence Changes in Time Spent with Media. *Journal of Broadcasting & Electronic Media* 60 (10 2016), 604–623. https://doi.org/10.1080/08838151.2016.1234477

[18] Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. 2019. SemEval-2019 Task 4: Hyperpartisan News Detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, Minneapolis, Minnesota, USA, 829–839. https://doi.org/10.18653/v1/S19-2145

[19] Klaus Krippendorff. 2011. Computing Krippendorff's Alpha-Reliability.

[20] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36, 4 (2020), 1234–1240.

[21] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 7871–7880. https://doi.org/10.18653/v1/2020.acl-main.703

[22] Sora Lim, Adam Jatowt, Michael Färber, and Masatoshi Yoshikawa. 2020. Annotating and Analyzing Biased Sentences in News Articles using Crowdsourcing. In *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 1478–1484. https://aclanthology.org/2020.lrec-1.184

[23] Sora Lim, Adam Jatowt, and Masatoshi Yoshikawa. 2018. Understanding Characteristics of Biased Sentences in News Articles. *CIKM Workshops*.

[24] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692 [cs.CL]

[25] Héctor Martínez Alonso, Amaury Delamaire, and Benoît Sagot. 2017. Annotating omission in statement pairs. In *Proceedings of the 11th Linguistic Annotation Workshop*. Association for Computational Linguistics, Valencia, Spain, 41–45. https://doi.org/10.18653/v1/W17-0805

[26] Malte Ostendorff, Till Blume, Terry Ruas, , Bela Gipp, and Georg Rehm. 2022. Specialized Document Embeddings for Aspect-based Similarity of Research Papers. In *2022 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. International Committee on Computational Linguistics, Kø", Germany (Online). https://doi.org/arXiv:2203.14541 Accepted for publication.

[27] Malte Ostendorff, Terry Ruas, Till Blume, Bela Gipp, and Georg Rehm. 2020. Aspect-based Document Similarity for Research Papers. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Barcelona, Spain (Online), 6194–6206. https://doi.org/10.18653/v1/2020.coling-main.545

[28] Malte Ostendorff, Terry Ruas, Moritz Schubotz, Georg Rehm, and Bela Gipp. 2020. Pairwise Multi-Class Document Classification for Semantic Relations between Wikipedia Articles. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*. ACM, Virtual Event China, 127–136. https://doi.org/10.1145/3383583.3398525

[29] Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2020. Automatically neutralizing subjective bias in text. In *Proceedings of the aaai conference on artificial intelligence*, Vol. 34. 480–489. https://doi.org/10.1609/aaai.v34i01.5385

[30] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. arXiv:1910.10683 [cs.LG]

[31] Marta Recasens and Dan Jurafsky. 2013. Linguistic Models for Analyzing and Detecting Biased Language. *ACL 2013 - 51st Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference* 1.

[32] Filipe N. Ribeiro, Lucas Henrique, Fabricio Benevenuto, Abhijnan Chakraborty, Juhi Kulshrestha, Mahmoudreza Babei, and Krishna P. Gummadi. 2015. Media Bias Monitor: Quantifying Biases of Social Media News Outlets at Large-Scale. In *Proceedings of the 12th International AAAI Conference of Web and Social Media* (Stanford, USA) *(ICWSM '18)*.

[33] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 4902–4912. https://doi.org/10.18653/v1/2020.acl-main.442

[34] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv:1910.01108 [cs.CL]

[35] Timo Spinde. 2021. An Interdisciplinary Approach for the Automated Detection and Visualization of Media Bias in News Articles. In *2021 IEEE International Conference on Data Mining Workshops (ICDMW)* (2021-09-30). https://doi.org/10.1109/ICDMW53433.2021.00144

[36] Timo Spinde, Felix Hamborg, Karsten Donnay, Angelica Becerra, and Bela Gipp. 2020. Enabling News Consumers to View and Understand Biased News Coverage: A Study on the Perception and Visualization of Media Bias. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020* (2020-01-01) *(JCDL '20)*. Association for Computing Machinery, Virtual Event, China, 389–392. https://doi.org/10.1145/3383583.3398619

[37] Timo Spinde, Felix Hamborg, and Bela Gipp. 2020. An Integrated Approach to Detect Media Bias in German News Articles. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020* (2020-01-01) *(JCDL '20)*. Association for Computing Machinery, Virtual Event, China, 505–506. https://doi.org/10.1145/3383583.3398585

[38] Timo Spinde, Felix Hamborg, and Bela Gipp. 2020. Media Bias in German News Articles : A Combined Approach. In *Proceedings of the 8th International Workshop on News Recommendation and Analytics (INRA 2020)* (2020-09-01). Virtual event. https://doi.org/10.1007/978-3-030-65965-3_41

[39] Timo Spinde, Christina Kreuter, Wolfgang Gaissmaier, Felix Hamborg, Bela Gipp, and Helge Giese. 2021. Do You Think It's Biased? How To Ask For The Perception Of Media Bias. arXiv:2112.07392 [cs.CL]

[40] Timo Spinde, David Krieger, Manu Plank, and Bela Gipp. 2021. Towards A Reliable Ground-Truth For Biased Language Detection. In *Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)* (2021-09-01). Virtual Event. https://media-bias-research.org/wp-content/uploads/2021/07/Spinde2021d_noDOI.pdf

[41] Timo Spinde, Jan-David Krieger, Terry Ruas, Jelena Mitrović, Franz Götz-Hahn, Akiko Aizawa, and Bela Gipp. 2022. Exploiting Transformer-based Multitask Learning for the Detection of Media Bias in News Articles. In *Proceedings of the iConference 2022* (2022-03-04). Virtual event. https://media-bias-research.org/wp-content/uploads/2021/11/Spinde2022a_mbg.pdf

[42] Timo Spinde, Manuel Plank, Jan-David Krieger, Terry Ruas, Bela Gipp, and Akiko Aizawa. 2021. Neural Media Bias Detection Using Distant Supervision With BABE - Bias Annotations By Experts. In *Findings of the Association for Computational Linguistics: EMNLP 2021* (2021-11-01). Dominican Republic. https://doi.org/10.18653/v1/2021.findings-emnlp.101

[43] Timo Spinde, Lada Rudnitckaia, Felix Hamborg, Bela, and Gipp. 2021. Identification of Biased Terms in News Articles by Comparison of Outlet-specific Word Embeddings. In *Proceedings of the iConference 2021* (2021-03-01). Beijing, China (Virtual Event). https://doi.org/10.1007/978-3-030-71305-8_17

[44] Timo Spinde, Lada Rudnitckaia, Sinha Kanishka, Felix Hamborg, Bela, Gipp, and Karsten Donnay. 2021. MBIC – A Media Bias Annotation Dataset Including Annotator Characteristics. In *Proceedings of the iConference 2021* (2021-03-01). Beijing, China (Virtual Event). https://media-bias-research.org/wp-content/uploads/2021/01/MBIC-\T1\textendash-A-Media-Bias-Annotation-Dataset-Including-Annotator-Characteristics.pdf

[45] Timo Spinde, Lada Rudnitckaia, Jelena Mitrović, Felix Hamborg, Michael Granitzer, Bela Gipp, and Karsten Donnay. 2021. Automated identification of bias inducing words in news articles using linguistic and context-oriented features. *Information Processing Management* 58, 3 (2021), 102505. https://doi.org/10.1016/j.ipm.2021.102505

[46] Timo Spinde, Kanishka Sinha, Norman Meuschke, and Bela Gipp. 2021. TASSY - A Text Annotation Survey System. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL)* (2021-09-01). https://doi.org/10.1109/JCDL52503.2021.00052

[47] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to Fine-Tune BERT for Text Classification? *CoRR* abs/1905.05583 (2019). arXiv:1905.05583 http://arxiv.org/abs/1905.05583

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[49] Jan Philip Wahle, Nischal Ashok, Terry Ruas, Norman Meuschke, Tirthankar Ghosal, and Bela Gipp. 2022. Testing the Generalization of Neural Language Models for COVID-19 Misinformation Detection. In *Information for a Better World: Shaping the Global Future*, Malte Smits (Ed.). Vol. 13192. Springer International Publishing, Cham, 381–392. https://doi.org/10.1007/978-3-030-96957-8_33 Series Title: Lecture Notes in Computer Science.

[50] Jan Philip Wahle, Terry Ruas, Tomáš Foltýnek, Norman Meuschke, and Bela Gipp. 2022. Identifying Machine-Paraphrased Plagiarism. In *Information for a Better World: Shaping the Global Future*, Malte Smits (Ed.). Vol. 13192. Springer International Publishing, Cham, 393–413. https://doi.org/10.1007/978-3-030-96957-8_34 Series Title: Lecture Notes in Computer Science.

[51] Jan Philip Wahle, Terry Ruas, Norman Meuschke, and Bela Gipp. 2021. Are Neural Language Models Good Plagiarists? A Benchmark for Neural Paraphrase Detection. In *2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE, Champaign, IL, USA, 226–229. https://doi.org/10.1109/JCDL52503.2021.00065 tex.ids= WahleRMG21 arXiv: 2103.12450.

[52] Jan Philip Wahle, Terry Ruas, Norman Meuschke, and Bela Gipp. 2021. Incorporating Word Sense Disambiguation in Neural Language Models. *arXiv:2106.07967 [cs]* (June 2021). https://arxiv.org/pdf/2106.07967.pdf arXiv: 2106.07967.

[53] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics, Brussels, Belgium, 353–355. https://doi.org/10.18653/v1/W18-5446

[54] Joseph Worsham and Jugal Kalita. 2020. Multi-task learning for natural language processing in the 2020s: Where are we going? *Pattern Recognition Letters* 136 (Aug. 2020), 120–126. https://doi.org/10/gmgb3j

[55] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*. 19–27. https://doi.org/10.1109/iccv.2015.11