Exploiting Transformer-based Multitask Learning for the Detection of Media Bias in News Articles

Timo Spinde[1] [0000-0003-3471-4127], Jan-David Krieger[2] [0000-0002-5360-2078], Terry Ruas[1] [0000-0002-9440-780X], Jelena Mitrović[3] [0000-0003-3220-8749], Franz Götz-Hahn[4] [0000-0003-3465-5040], Akiko Aizawa[5] [0000-0001-6544-5076], and Bela Gipp[1] [0000-0001-6522-3019]

[1] University of Wuppertal, Germany
{firstname.lastname}@uni-wuppertal.de
[2] University of Konstanz, Germany
Jan-David.Krieger@uni-konstanz.de
[3] University of Passau, Germany
Institute for Artificial Intelligence Research and Development of Serbia
jelena.mitrovic@Uni-Passau.de
[4] University of Kassel, Germany
franz.goetz-hahn@uni-kassel.de
[5] NII Tokyo, Japan
aizawa@nii.ac.jp

Abstract. Media has a substantial impact on the public perception of events. A one-sided or polarizing perspective on any topic is usually described as media bias. One of the ways how bias in news articles can be introduced is by altering word choice. Biased word choices are not always obvious, nor do they exhibit high context-dependency. Hence, detecting bias is often difficult. We propose a Transformer-based deep learning architecture trained via Multi-Task Learning using six bias-related data sets to tackle the media bias detection problem. Our best-performing implementation achieves a macro F_1 of 0.776, a performance boost of 3% compared to our baseline, outperforming existing methods. Our results indicate Multi-Task Learning as a promising alternative to improve existing baseline models in identifying slanted reporting.

Keywords: Media Bias · Text Analysis · Multi-Task Learning · News Analysis

1 Introduction

Media bias, i.e., slanted news coverage, has the potential to drastically change the public opinion on any topic [32]. One of the forms bias can be expressed by is by word choice, e.g., depicting any content in a non-neutral way [23]. Detecting and highlighting media bias might be relevant for media analysis and mitigate the effects of biased reporting. To date, only a few research projects focus on the detection and aggregation of bias [6, 16]. One of the reasons that make the creation of automated methods to detect media bias a complex task is often the subtle nature of media bias, which represents a challenge for quantitative identification methods [10, 16, 30, 33]. While many current research projects focus on collecting linguistic features to describe media bias [11, 23, 29, 34], we propose a Transformer-based [39] architecture for the classification of media bias. Similar models have recently shown to achieve performance increases in the media bias domain, e.g., sentence-level bias detection [6, 12, 27, 32]. However, so far, they rely on very limited resources. Data sets with bias gold standard annotations are, to date, only scarcely available, and exhibit various weaknesses, such as low inter annotator agreement, small size, or no information about the annotator background [31, 32, 36]. Additionally, state-of-the-art neural language models usually require large amounts of training data to yield meaningful representations [9, 24], which are incompatible with the size of current media bias data sets [10, 35]. To mitigate the lack of suitable data sets, our model incorporates Multi-Task Learning (MTL) [24], which allows for increasing performance by sharing model representations between related tasks [13, 19, 37]. The use of cross-domain data sets in our model is particularly relevant for the media bias domain as multiple sources can provide a more robust model. To the best of our knowledge, the MTL paradigm has not been explored in existing work on media bias. Our research question is therefore to assess whether MTL can improve models to classify media bias automatically.

The main contribution of this paper is to incorporate Transformer-based MTL into a system to identify sentence-level media bias automatically. We exploit MTL in the media bias context by computing multiple models based on different combinations of auxiliary training data sets (section 2). All our models, data, and code are publicly available on https://bit.ly/3cmiQgB.

2 Related Work

While there are multiple forms of media bias, e.g., bias by personal perception or by the omission of information [28], our focus is on bias by word choice, in which different words refer to the same concept [23]. We will first summarize available media bias data sets and then present automated methods to identify bias as well as MTL.

The concept of media bias is covered by many data sets [1, 6, 10, 18, 35]. However, they all exhibit specific deficiencies, such as (1) a low number of topics [16, 18], (2) no annotations on the word level [18], (3) low inter-annotator agreement [1,17,18,35], (4) no background check for its participants (except [35]), and (5) only article-level annotations [6]. Also, some related papers focus on framing rather than on bias [1, 10], or on Wikipedia instead of news [12], and results are only partially transferable. To the best of our knowledge, the most extensive and precise data set was presented recently [32]. The data set consists of 3700 sentences annotated by expert raters on sentence-level with an inter-annotator agreement of 0.40 measured by Krippendorff's α [15], which is higher than for all other available data sets.

Several studies tackle the challenge of identifying media bias automatically [6, 11, 12, 23, 34]. Most of them use hand-crafted features to detect bias [11, 34]. For example, [34] identify and evaluate a wide range of linguistic, lexical, and syntactic features serving as potential bias indicators. The existing work on neural models is based on the data sets mentioned above, which exhibit the described weaknesses [6, 12]. Most media bias models focus on sentence-level bias [6, 10–12, 23]. Therefore, we follow the standard practice and construct a sentence-level classifier.

MTL approaches have shown to be helpful when high-quality data sets in the domain are scarce, but text corpora covering general related concepts are available [13,19,37,38,40]. For example, [13] report that MTL applied on BERT yields an accuracy increase of 1.03% compared to the baseline BERT in a subjectivity detection task. MTL might be a suitable training paradigm for media bias identification systems since sufficiently sized bias corpora with qualitative handcrafted annotations do not exist. Therefore, we propose the first neural MTL media bias classifier composed of inter-domain and cross-domain data sets.

3 Methodology

We explore how fine-tuning a language model via MTL can improve the performance in detecting media bias on the sentence level. Computational costs are an important consideration for us since we train multiple large-scale MTL models. For this reason, we employ a distilled modification of BERT [9], called DistilBERT [26], which achieves a 40% reduction in size while simultaneously accelerating the training process by 60% and retaining 97% of language understanding capabilities on NLP benchmark tasks [40]. DistilBERT represents an appropriate architecture, keeping resource consumption and performance balanced. The incorporation of larger models trained via MTL is left to future research.

Our MTL technique is based on *hard parameter sharing* in which all hidden model layers are shared between auxiliary training tasks [24]. Task-specific layers are added on top of the last hidden state, accounting for the label structure of auxiliary data sets. The MTL paradigm we propose is architecture-independent and can be adjusted to future NLP architectures.

For our training procedure, we distinguish between models trained on in-domain and cross-domain data sets. For in-domain data sets, the creation process included concepts related to media bias, such as subjectivity [21]. Conversely, cross-domain data sets include data points that are not directly annotated for or related to media bias, but are retrieved from tasks that bear some connection to it. The auxiliary data sets we use comprise a diverse set of NLP tasks requiring two different losses for the learning process – the Cross-Entropy (CE) loss [8] and the Mean Squared Error (MSE) loss [25]. The origin and number of the data used for the training of our models, as well as their respective original tasks and used loss functions, are shown in Table 1. We use in-domain (ID) and cross-domain (CD) data sets used in other MTL studies within the language processing domain [13, 19, 37].

Data set	**Domain**	n	**Task**	**Loss**	**Description**
Reddit data set (Reddit) [4]	ID	6861	Single Sentence Regression	MSE	Reddit comments labeled on a continuous scale ranging from 0 (supportive) to 1 (discriminatory)
Subjectivity data set (Subj) [21]	ID	10000	Single Sentence Classification	CE	Movie reviews labeled as *objective* or *opinionated*
IMDb [20]	ID	50000	Single Sentence Classification	CE	Movie reviews containing positive and negative sentiment labels
Wikipedia data set (Wiki)[1] [22]	ID	180000	Single Sentence Classification	CE	Neutral and biased sentence pairs from articles going against Wikipedia's NPOV rule
Semantic Textual Similarity Benchmark (STS-B) data set [5]	CD	10943	Pairwise Sentence Similarity	MSE	Multilingual and cross-lingual sentence pairs labeled in terms of similarity
Stanford Natural Language Inference (SNLI) corpus [1] [2]	CD	570000	Pairwise Sentence Classification	CE	Sentence pairs labeled for linguistic relations within the labels *entailment*, *neutral*, or *contradiction*

[1] We only use 50000 text instances from these corpora in our MTL approach to keep the size of training sets balanced.

Table 1. Auxiliary data sets incorporated in the MTL models (n = number of instances)

fig. 1 outlines our in-domain MTL model consisting of DistilBERT's encoder, whose parameters are shared across tasks, and the added task-specific layers [6]. The represented model is based on the maximum number of possible data sets within the approach. In our experiments on MTL, we try various combinations, including at least three in-domain and five cross-domain data sets, respectively.

For preprocessing and MTL training, we took the same approach as [19]. Initially, pre-trained parameters are loaded from *huggingface* [7]. We split up data for a fixed-size subset of tasks into batches, and batches are merged and shuffled to guarantee the model does not train on too many subsequent batches of a single task. The preprocessing step is repeated every epoch. Batches are then passed on by the data loader one by one to the model, which outputs task-specific predictions and the respective loss. Finally, the loss is backpropagated, and parameters are updated.

4 Experiments

To investigate the benefit of MTL to identify media bias on a fine-grained linguistic level, we train ten models using MTL, which we compare to five baseline models. As a consequence of a lack of robust guidelines for selection criteria for

[6] The cross-domain model is not shown due to lack of space but is published at the repository mentioned in Section section 1.

[7] https://huggingface.co/transformers/model_doc/distilbert.html

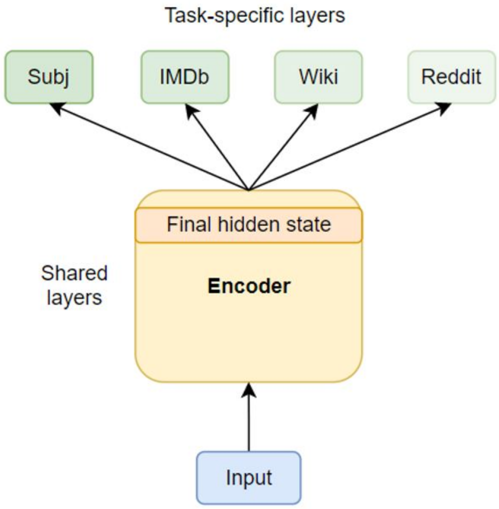

Fig. 1. Outline of in-domain MTL model consisting of a shared encoder block and task-specific layers. *Note*: We implement multiple MTL models based on different combinations of the presented data sets.

auxiliary corpora, we choose a variety of auxiliary tasks to fine-tune the DistilBERT model via MTL that have previously been used successfully in MTL studies [13,19,37]. Each of our MTL models is trained using a different combination of a sample of six popular data sets, where IMDb [20], Subj [21], Wiki [22], and Reddit [4] are considered in-domain data sets, and STS-B [5], and SNLI [2] comprise examples of cross-domain data sets[8].

The in-domain models are based on bias-related data sets [9]. Combining the in-domain corpora yields five different models (table 2, M1 - M5): four use triple combinations, and one model relies on all in-domain data sets. The cross-domain models extend the pool of experiments by adding the STS-B and SNLI data sets to each of the five in-domain models (table 2, M6 - M10). The approaches are oriented on the MTL fine-tuning approaches applied in [37]. In their experiments based on BERT, the authors apply MTL on domain-related and domain-unrelated data yielding a performance boost for sentiment classification.

All experiments are performed on a *Google Colab NVIDIA Tesla K80* [10]. We choose the *AdamW optimizer* [14] and a batch size of 32 . All downstream task layers are based on a hidden state dimensionality of 768. All performance metrics are calculated based on 5-fold cross-validation [3]. Thus, we divide the

[8] A detailed description of the data sets is published at the repository mentioned in Section section 1.

[9] IMDb, Subj, Wiki, Reddit

[10] https://colab.research.google.com/notebooks/intro.ipynb

final bias data set containing 1700 instances into five different train and tests [11]. The models are then iteratively trained on all five training sets and evaluated on the respective held-out test set. Finally, the performance metrics on the test sets are averaged, yielding the cross-validated model performance. Each respective model is trained over four epochs with an early stopping criterion based on validation CE loss. In many cases, the model stops learning after two epochs. The MTL fine-tuning is based on a learning rate of $5 \cdot 10^{-5}$.

As far as we know, there are no related works applying MTL in the media bias domain. Therefore, we compare the performances of our MTL approaches to a set of baseline models (table 2; B1-B5). We report the performance scores achieved from pre-trained DistilBERT provided by *huggingface* (B1). Furthermore, we train four DistilBERT models on each of the in-domain data sets (B2-B5). Thus, we can observe whether the assumed performance boost of our MTL models results from MTL rather than domain-relatedness of the training data.

We expect that fine-tuning via MTL leads to an improvement of DistilBERT's bias identification power. Mainly, we want to analyze whether the MTL technique yields a substantial performance boost compared to simple Transfer Learning (TL) approaches training the model on only a single data set. Therefore, we run several experiments.

5 Results and Discussion

We show the performance indicators of our model on our expert-labeled media bias data set in table 2, according to F_1, precision, recall, and loss. Since the highest macro F_1 score does not necessarily match with the lowest loss, we elaborate on the results from the perspective of both metrics.

Among all MTL-trained models the highest F_1 score is achieved from the in-domain M4 model with 0.776. The best cross-domain model regarding macro F_1 is reached by M8 with 0.771. Compared to DistilBERT, M4 achieves a 3% increase in macro F_1, while B5 achieves the highest macro F_1 for TL-based models at 0.782, which is not surpassed by any MTL approach. Although all MTL models outperform DistilBERT, the highest macro F_1 score of all MTL models is 0.6% lower than that of B5. Overall, MTL improves the B1 baseline macro F_1 score in a range from 0.3% (M9) to 3% (M4). When considering the models from a loss-based perspective, the performance ranks change slightly: M4 remains as the best in-domain MTL model, but M7 (the second to last in terms of macro F_1 performance) reaches the lowest loss within the cross-domain approaches. Compared to DistilBERT, M4 shows a decrease in loss of 4.9%. B5 prevails as the best TL model with a CE loss of 0.466. In contrast to the macro F_1-based perspective, however, M4 achieves the lowest overall loss, outperforming B5 by 0.2%.

In general, our MTL approaches surpass the baseline methods. However, the best overall model based on macro F_1 was a TL model trained on a data

[11] We use a subset of BABE [32], introduced in section 2, to evaluate the MTL models.

	Model	Data sets						macro F_1	micro F_1	binary F_1	Precision	Recall	CE Loss
	B1	*huggingface* DistilBERT						0.746	0.750	0.711	**0.805**	0.640	0.513
TL	B2	✓						0.744	0.744	0.730	0.744	0.716	0.545
	B3		✓					0.761	0.762	0.746	0.770	0.725	0.491
	B4			✓				0.743	0.746	0.709	0.790	0.646	0.497
	B5				✓			**0.782**	**0.782**	**0.7695**	0.785	0.754	0.466
ID MTL	M1	✓	✓	✓				0.768	0.768	0.753	0.778	0.731	0.482
	M2	✓	✓		✓			0.760	0.760	0.746	0.766	0.729	0.495
	M3	✓		✓	✓			0.773	0.774	0.762	0.777	0.755	0.482
	M4		✓	✓	✓			0.776	0.777	0.759	0.794	0.727	**0.464**
	M5	✓	✓	✓	✓			0.772	0.771	0.757	0.778	0.737	0.473
CD MTL	M6	✓	✓	✓		✓	✓	0.766	0.766	0.758	0.756	0.763	0.492
	M7	✓	✓		✓	✓	✓	0.765	0.765	0.751	0.770	0.735	0.474
	M8	✓		✓	✓	✓	✓	0.771	0.771	0.762	0.765	0.761	0.491
	M9		✓	✓	✓	✓	✓	0.749	0.750	0.759	0.714	**0.812**	0.499
	M10	✓	✓	✓	✓	✓	✓	0.769	0.770	0.751	0.789	0.720	0.480
		Subj	IMDb	Reddit	Wiki	STS	SNLI						

Table 2. Results for all baseline models, i.e., the *huggingface* model or models obtained by TL, as well as the models trained using MTL considering only in-domain data sets or also incorporating cross-domain data. For each metric we have denoted the best performance in bold.

set containing revised Wikipedia excerpts (B5). Based on CE loss, only one MTL model slightly outperform this TL model. Thus, we cannot state whether Transformer-Based MTL improves media bias detection on the sentence level. We assume that the strong performance of B5 results from the relatedness of the underlying data set to our bias corpus. The Wikipedia data set contains biased and neutral sentences extracted from revised Wikipedia passages. Hence, the data set is similar to our bias corpus[12]. The only difference to our fine-tuning data set is the source from which the data is extracted. Pre-training a Transformer-based model on a highly bias-related corpus seems to hinder MTL's relative performance improvement. Furthermore, we assume that our selection of auxiliary data sets might not have been sufficiently comprehensive. In our MTL approaches, updating DistilBERT's parameters only required the computation and back-propagation of binary CE loss and MSE loss. [24] argues that well-performing MTL approaches must be trained on NLP tasks, including multiple loss functions.

Existing MTL studies [13, 19, 37] do not report diverse TL baseline models. The MTL approaches are primarily compared to a pre-trained baseline model provided by model libraries. Future research should incorporate a comprehensive set of baseline models allowing for a more robust analysis. Comparing our best MTL model to DistilBERT, the effect of MTL is similar as in [13].

Considering our MTL-based media bias research, future work should include more comprehensive sets of bias-related auxiliary data sets with multiple loss functions. Possible tasks could, for example, comprise the detection of bias-inducing linguistic features such as negative sentiment [34]. In this way, deep

[12] Let us point out that **none** of the instances from the Wikipedia data set are contained in our target media bias data set.

learning techniques could benefit from other types of tasks, such as classifying linguistic features. Moreover, future MTL approaches could benefit from larger transformer models (e.g., XLNet [41], ELECTRA [7]). Our approach based on DistilBERT is the first step towards balancing cloud-computing costs and performance. We note that a follow-up experiment about an improved model and a larger exploratory data analysis are already in progress and will be published in the future.

6 Final Considerations

This work proposes a Transformer-based MTL approach to identify media bias by word choice in news articles. The motivation for selecting the training technique results from our observation that the size of available media bias data sets is not compatible with the requirements of state-of-the-art neural language models. We train ten MTL models based on different combinations of six auxiliary data sets and compare them to five baseline models. Our results show that the best performing MTL model partly surpasses the baseline models in terms of macro F_1 loss and CE loss. Yet, we can not ascertain a significant superiority of the MTL approach in classifying media bias instances. The main limitation of our work is the restricted inclusion of auxiliary tasks. In future work, we plan to incorporate more tasks based on bias-inducing linguistic features. We have to emphasize that any successful MTL implementation in the context of media bias identification could decrease financial burdens emerging from the collection of hand-crafted training data. Yet, at the same time, cloud computing requires substantial financial resources. Costs of using larger models should therefore be properly evaluated. We believe the MTL approach to be promising in the area and aim to continue the research on MTL in connection with media bias identification in the future.

References

1. Baumer, E., Elovic, E., Qin, Y., Polletta, F., Gay, G.: Testing and comparing computational approaches for identifying the language of framing in political news. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 1472–1482. Association for Computational Linguistics, Denver, Colorado (May–Jun 2015). https://doi.org/10.3115/v1/N15-1171, https://www.aclweb.org/anthology/N15-1171
2. Bowman, S.R., Angeli, G., Potts, C., Manning, C.D.: A large annotated corpus for learning natural language inference. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 632–642. Association for Computational Linguistics, Lisbon, Portugal (Sep 2015). https://doi.org/10.18653/v1/D15-1075, https://www.aclweb.org/anthology/D15-1075
3. Browne, M.W.: Cross-validation methods. J. Math. Psychol. **44**(1), 108–132 (Mar 2000). https://doi.org/10.1006/jmps.1999.1279, https://doi.org/10.1006/jmps.1999.1279

4. Cabot, P.H., Abadi, D., Fischer, A., Shutova, E.: Us vs. them: A dataset of populist attitudes, news bias and emotions. CoRR **abs/2101.11956** (2021), https://arxiv.org/abs/2101.11956
5. Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., Specia, L.: Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017) (2017). https://doi.org/10.18653/v1/s17-2001, http://dx.doi.org/10.18653/v1/S17-2001
6. Chen, W.F., Al Khatib, K., Wachsmuth, H., Stein, B.: Analyzing Political Bias and Unfairness in News Articles at Different Levels of Granularity. In: Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science. pp. 149–154. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.nlpcss-1.16, https://www.aclweb.org/anthology/2020.nlpcss-1.16
7. Clark, K., Luong, M.T., Le, Q.V., Manning, C.D.: ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. arXiv:2003.10555 [cs] (Mar 2020)
8. De Boer, P.T., Kroese, D.P., Mannor, S., Rubinstein, R.Y.: A tutorial on the cross-entropy method. Annals of operations research **134**(1), 19–67 (2005)
9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: 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). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1423, https://www.aclweb.org/anthology/N19-1423
10. Fan, L., White, M., Sharma, E., Su, R., Choubey, P.K., Huang, R., Wang, L.: In plain sight: Media bias through the lens of factual reporting. 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). pp. 6343–6349. Association for Computational Linguistics, Hong Kong, China (Nov 2019). https://doi.org/10.18653/v1/D19-1664, https://www.aclweb.org/anthology/D19-1664
11. Hube, C., Fetahu, B.: Detecting biased statements in wikipedia. In: Companion Proceedings of the The Web Conference 2018. p. 1779–1786. WWW '18, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE (2018). https://doi.org/10.1145/3184558.3191640, https://doi.org/10.1145/3184558.3191640
12. Hube, C., Fetahu, B.: Neural based statement classification for biased language. In: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining. p. 195–203. WSDM '19, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3289600.3291018, https://doi.org/10.1145/3289600.3291018
13. Huo, H., Iwaihara, M.: Utilizing bert pretrained models with various fine-tune methods for subjectivity detection. In: Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data. pp. 270–284. Springer (2020)
14. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations,

ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015), http://arxiv.org/abs/1412.6980

15. Krippendorff, K.: Computing krippendorff's alpha-reliability. Departmental Papers (ASC); University of Pennsylvania (2011), https://repository.upenn.edu/cgi/viewcontent.cgi?article=1043&context=asc_papers
16. Lim, S., Jatowt, A., Färber, M., Yoshikawa, M.: Annotating and analyzing biased sentences in news articles using crowdsourcing. In: Proceedings of the 12th Language Resources and Evaluation Conference. pp. 1478–1484. European Language Resources Association, Marseille, France (May 2020), https://www.aclweb.org/anthology/2020.lrec-1.184
17. Lim, S., Jatowt, A., Färber, M., Yoshikawa, M.: Annotating and analyzing biased sentences in news articles using crowdsourcing. In: Proceedings of the 12th Language Resources and Evaluation Conference. pp. 1478–1484. European Language Resources Association, Marseille, France (May 2020), https://www.aclweb.org/anthology/2020.lrec-1.184
18. Lim, Sora and Jatowt, Adam and Yoshikawa, Masatoshi: Understanding Characteristics of Biased Sentences in News Articles. In: CIKM Workshops (2018), {http://ceur-ws.org/Vol-2482/paper13.pdf}
19. Liu, X., He, P., Chen, W., Gao, J.: Multi-task deep neural networks for natural language understanding. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 4487–4496. Association for Computational Linguistics, Florence, Italy (Jul 2019), https://www.aclweb.org/anthology/P19-1441
20. Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y., Potts, C.: Learning word vectors for sentiment analysis. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. pp. 142–150. Association for Computational Linguistics, Portland, Oregon, USA (Jun 2011), https://www.aclweb.org/anthology/P11-1015
21. Pang, B., Lee, L.: A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. p. 271–es. ACL '04, Association for Computational Linguistics, USA (2004). https://doi.org/10.3115/1218955.1218990, https://doi.org/10.3115/1218955.1218990
22. Pryzant, R., Diehl Martinez, R., Dass, N., Kurohashi, S., Jurafsky, D., Yang, D.: Automatically neutralizing subjective bias in text. Proceedings of the AAAI Conference on Artificial Intelligence **34**(01), 480–489 (Apr 2020). https://doi.org/10.1609/aaai.v34i01.5385, https://ojs.aaai.org/index.php/AAAI/article/view/5385
23. Recasens, M., Danescu-Niculescu-Mizil, C., Jurafsky, D.: Linguistic models for analyzing and detecting biased language. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1650–1659 (2013), https://www.aclweb.org/anthology/P13-1162.pdf
24. Ruder, S.: An overview of multi-task learning in deep neural networks. CoRR **abs/1706.05098** (2017), http://arxiv.org/abs/1706.05098
25. Sammut, C., Webb, G.I. (eds.): Mean Squared Error, pp. 653–653. Springer US, Boston, MA (2010). https://doi.org/10.1007/978-0-387-30164-8_528, \url{https://doi.org/10.1007/978-0-387-30164-8_528}
26. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. CoRR **abs/1910.01108** (2019), http://arxiv.org/abs/1910.01108

27. Spinde, T.: 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), https://media-bias-research.org/wp-content/uploads/2021/09/Spinde2021g.pdf
28. Spinde, T., Hamborg, F., Donnay, K., Becerra, A., Gipp, B.: 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. p. 389–392. JCDL '20, Association for Computing Machinery, Virtual Event, China (2020). https://doi.org/10.1145/3383583.3398619, https://doi.org/10.1145/3383583.3398619
29. Spinde, T., Hamborg, F., Gipp, B.: Media bias in german news articles: A combined approach. ECML PKDD 2020 Workshops: Workshops of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2020): INRA 2020, Ghent, Belgium, September 14–18, 2020, Proceedings **1323**, 581–590 (2020). https://doi.org/10.1007/978-3-030-65965-3_41, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7850083/
30. Spinde, T., Kreuter, C., Gaissmaier, W., Hamborg, F., Gipp, B., Giese, H.: 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)
31. Spinde, T., Krieger, D., Plank, M., Gipp, B.: Towards A Reliable Ground-Truth For Biased Language Detection. In: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL) (Sep 2021)
32. Spinde, T., Plank, M., Krieger, J.D., Ruas, T., Gipp, B., Aizawa, A.: Neural Media Bias Detection Using Distant Supervision With BABE - Bias Annotations By Experts. In: Findings of the Association for Computational Linguistics: EMNLP 2021. Dominican Republic (Nov 2021)
33. Spinde, T., Rudnitckaia, L., Hamborg, F., Gipp, B.: Identification of biased terms in news articles by comparison of outlet-specific word embeddings. In: Proceedings of the iConference 2021 (March 2021)
34. Spinde, T., Rudnitckaia, L., Mitrović, J., Hamborg, F., Granitzer, M., Gipp, B., Donnay, K.: Automated identification of bias inducing words in news articles using linguistic and context-oriented features. Information Processing & Management **58**(3), 102505 (2021). https://doi.org/10.1016/j.ipm.2021.102505, https://doi.org/10.1016/j.ipm.2021.102505
35. Spinde, T., Rudnitckaia, L., Sinha, K., Hamborg, F., Gipp, B., Donnay, K.: MBIC – A media bias annotation dataset including annotator characteristics. In: Proceedings of the iConference 2021. iSchools (2021). https://doi.org/10.5281/zenodo.4474336
36. Spinde, T., Sinha, K., Meuschke, N., Gipp, B.: TASSY - A Text Annotation Survey System. In: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL) (Sep 2021)
37. Sun, C., Qiu, X., Xu, Y., Huang, X.: How to fine-tune bert for text classification? In: Sun, M., Huang, X., Ji, H., Liu, Z., Liu, Y. (eds.) Chinese Computational Linguistics. pp. 194–206. Springer International Publishing, Cham (2019)
38. Sun, Y., Wang, S., Li, Y.K., Feng, S., Tian, H., Wu, H., Wang, H.: Ernie 2.0: A continual pre-training framework for language understanding. Proceedings of the AAAI Conference on Artificial Intelligence **34**, 8968–8975 (04 2020). https://doi.org/10.1609/aaai.v34i05.6428

39. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017), https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
40. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.R.: Glue: A multitask benchmark and analysis platform for natural language understanding. arXiv preprint arXiv:1804.07461 (2018)
41. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., Le, Q.V.: XLNet: Generalized Autoregressive Pretraining for Language Understanding. arXiv:1906.08237 [cs] (Jun 2019)