MAGPIE: Multi-Task Analysis of Media-Bias Generalization with Pre-Trained Identification of Expressions

Tomáš Horych[1,3,6], **Martin Wessel**[4,6], **Jan Philip Wahle**[2], **Terry Ruas**[2], **Jerome Waßmuth**[5], **André Greiner-Petter**[2,3], **Akiko Aizawa**[3], **Bela Gipp**[2], **Timo Spinde**[2,6]

[1]Czech Technical University, Prague, Czech Republic
[2]University of Göttingen, Göttingen, Germany, {*last*}@uni-goettingen.de
[3]National Institute of Informatics, Tokyo, Japan, {*last*}@nii.ac.jp
[4]CDTM, TU Munich, Germany, [5]University of Konstanz, Germany, {*last*}@uni-konstanz.de
[6]{t.horych, m.wessel, t.spinde}@media-bias-research.org

Abstract

Media bias detection poses a complex, multifaceted problem traditionally tackled using single-task models and small in-domain datasets, consequently lacking generalizability. To address this, we introduce MAGPIE, a large-scale multi-task pre-training approach explicitly tailored for media bias detection. To enable large-scale pre-training, we construct Large Bias Mixture (LBM), a compilation of 59 bias-related tasks. MAGPIE outperforms previous approaches in media bias detection on the Bias Annotation By Experts (BABE) dataset, with a relative improvement of 3.3% F1-score. Furthermore, using a RoBERTa encoder, we show that MAGPIE needs only 15% of fine-tuning steps compared to single-task approaches. We provide insight into task learning interference and show that sentiment analysis and emotion detection help learning of all other tasks, and scaling the number of tasks leads to the best results. MAGPIE confirms that MTL is a promising approach for addressing media bias detection, enhancing the accuracy and efficiency of existing models. Furthermore, LBM is the first available resource collection focused on media bias MTL.

Keywords: Media bias, Multi-task learning, Text classification

1. Introduction

Media bias is a skewed portrayal of information favoring certain group interests (Recasens et al., 2013), which manifests in multiple facets, including political, gender, racial, and linguistic biases. Such subtypes of bias, which can intersect and coexist in complex combinations, make the classification of media bias a challenging task (Raza et al., 2022). Existing research on media bias detection primarily involves training classifiers on small in-domain datasets (Krieger et al., 2022), which exhibit limited generalizability across diverse domains (Wessel et al., 2023).

This paper builds upon the work of Wessel et al. (2023), emphasizing that the multifaceted nature of media bias detection requires a shift from isolated approaches to multi-task methodologies, considering a broad spectrum of bias types and datasets. The recent advancements in Multi-Task Learning (MTL) (Aribandi et al., 2021a; Chen et al., 2021; Kirstein et al., 2022) open up promising opportunities to overcome these challenges by enabling knowledge transfer across domains and tasks. Despite the potential, a comprehensive MTL approach for media bias detection is yet to be realized. The only other media bias MTL method (Spinde et al., 2022) underperforms due to its narrow task focus and does not surpass baseline outcomes.

In this study, we make five main contributions:

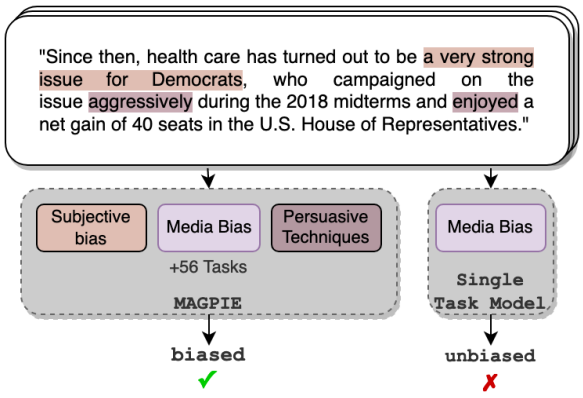


Figure 1: MAGPIE has a pre-trained representation of multiple biases (persuasive, subjective, etc.). This enables it to outperform models based on single-task learning (STL) paradigms.

1. We present **MAGPIE** - the first large-scale multi-task pre-training approach for media bias detection. By pre-training on diverse bias types such as persuasive and subjective, a classifier based on MAGPIE correctly classifies sentences that state-of-the-art single-task models misidentify (we show an example in Figure 1).

LREC-COLING 2024, pages 10903–10920
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

2. We introduce **LBM** (Large Bias Mixture), a pre-training composition of 59 bias-related tasks encompassing wide range of biases such as linguistic bias, gender bias and group bias.

3. We provide an analysis of a task selection and demonstrate the effectiveness of scaling the number of tasks.

4. We demonstrate that MAGPIE outperforms the previous state-of-the-art model by 3.3% on the Media Bias Annotation by Experts (BABE) dataset (Spinde et al., 2021c) and achieves competetive results on the Media Bias Identification Benchmark (MBIB) collection (Wessel et al., 2023).

5. We make all resources, including datasets, training framework, documentation, and models, publicly available on GitHub:

github.com/magpie-multi-task

These contributions highlight the potential of MTL in improving media bias detection. Our findings show, e.g., that tasks like sentiment and emotionality enhance overall learning, all tasks boost fake news detection, and scaling tasks leads to optimal results. Another key insight of our research is the value of MTL in contexts where the primary dataset is small[1]. By learning generalized bias knowledge from a range of tasks, we can improve the accuracy and efficiency of existing models, even in the face of limited data. Overall, our research offers a multi-task learning approach to media bias detection with first large-scale resources in the domain.

2. Related Work

2.1. Media Bias

Media bias is a complex issue (Lee et al., 2021a; Recasens et al., 2013; Raza et al., 2022) composed of varying definitions of bias subtypes such as linguistic bias, context bias, or group bias (Wessel et al., 2023). In their literature review, Spinde et al. (2023) provide an extensive overview of research on media bias and related subtypes of bias.

Media bias detection approaches have evolved from hand-crafted features (Recasens et al., 2013; Hube and Fetahu, 2018; Spinde et al., 2021d) to neural models (Spinde et al., 2022; Chen et al., 2021; Spinde et al., 2021c; Huguet Cabot et al., 2021; Sinha and Dasgupta, 2021; Raza et al., 2022). However, existing models, so far, focus only on single tasks and saturate quickly on smaller datasets (Wessel et al., 2023). As most neural approaches require large quantities of data, those relying on single and small datasets cannot provide a realistic scenario for their solutions (e.g., Fan et al. (2019)). We will first provide an overview of existing datasets and then show how to exploit their diversity within the media bias domain.

Media bias tasks and datasets mainly cover individual, self-contained tasks such as binary classifications (Recasens et al., 2013; Spinde et al., 2021b), which, so far, are not explored in relation to each other (Spinde et al., 2023). Wessel et al. (2023) systematically form the media bias detection benchmark MBIB by reviewing over 100 media bias datasets and consolidating 22 of them into eight distinct tasks like linguistic, racial, and political bias. Their study highlights that methods only focused on one of these tasks exhibit limitations in their detection capabilities. MAGPIE encompasses all the tasks identified in the MBIB but also significantly expands its scope by incorporating an additional 51 media bias-related tasks to mitigate a variety of limitations in MBIB (see Section 3.1).

2.2. Multi-Task Learning

MTL shows significant improvements in various NLP tasks, including sentiment analysis (He et al., 2019), text summarization (Kirstein et al., 2022), and natural language understanding (Raffel et al., 2020). In MTL, a model leverages knowledge gained from one task to improve the performance of others. Aribandi et al. (2021b) demonstrate that increasing the number of tasks generally leads to improved performance for downstream NLP applications. Aghajanyan et al. (2021) show that pre-finetuning, a large-scale multi-task learning phase, consistently improves the performance and efficiency of pre-trained models across diverse tasks, with results improving linearly with the number of tasks beyond a certain threshold.

As described above, media bias can be seen as a composite problem composed of various interrelated bias types (Wessel et al., 2023). In the realm of Natural Language Understanding (NLU), MTL has proven to be highly effective when incorporating related tasks Aribandi et al. (2021b). For instance, benchmarks such as GLUE and SuperGLUE successfully decompose the NLU problem into a suite of proxy tasks, including paraphrase detection (Wahle et al., a,b), and semantic evaluation (Ruas et al., b,a), thereby substantially improving performance across a range of NLU tasks (Wang et al., 2018, 2019). Motivated by this success in NLU, we propose to jointly learn from different bias types within the media bias domain. With this approach, we aim to treat media bias not as a singular entity but as many interconnected issues.

The selection of tasks is pivotal to the efficacy of MTL. There have been several attempts to au-

[1] For example, the Media Bias Annotation by Experts (BABE) dataset (Spinde et al., 2021c).

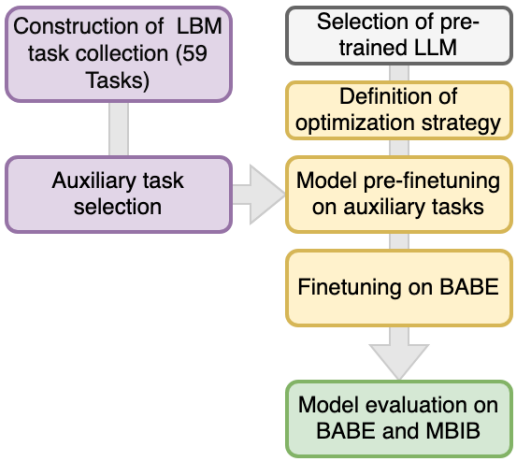


Figure 2: The process of training and evaluating MAGPIE. The purple steps describe the construction and usage of LBM, the yellow the model training, and the green the model evaluation.

tomate task selection, including learning the data selection with Bayesian optimization (Bingel and Søgaard, 2017) or estimating task relations (Ruder and Plank, 2017). The most model-agnostic approach is GradTS (Ma et al., 2021), which is highly scalable due to low resource requirements, and is therefore implemented within MAGPIE. GradTS accumulates gradients of attention heads and selects tasks based on their correlation with the primary task's attention. The selected tasks are trained jointly and share representations across tasks.

3. Methodology

We implement MAGPIE using pre-finetuning as introduced in (Aghajanyan et al., 2021) (See also Section 2). As such, MAGPIE is an encoder-only MTL transformer model pre-finetuned on 59 media bias-related auxiliary tasks provided by Large Bias Mixture (LBM), a large-scale task collection of bias-related datasets. We incorporate a novel approach of a Head-Specific Early Stopping and Resurrection to effectively handle tasks of varying sizes (Section 3.3.1).

As outlined in Figure 2, our first step involves constructing the LBM. Following this, we define the model and multi-task learning (MTL) framework employed to train MAGPIE, which includes optimization strategies, task sampling, and auxiliary task selection. Lastly, we evaluate MAGPIE on two primary resources: the Media Bias Annotation by Experts (BABE) dataset[2] (Spinde et al., 2021c), and the Media Bias Identification Benchmark (MBIB) collection.

3.1. The LBM (Large Bias Mixture) task collection

Currently, MBIB is the only collection of media bias tasks. However, it does not include tasks that constitute a form of media bias more indirectly, such as sentiment analysis or emotion classification. MAGPIE aims to integrate tasks both directly linked to media bias and those peripherally related, such as sentiment analysis, to provide broader coverage of linguistic features in the media bias context. Therefore, we introduce Large Bias Mixture (LBM), a more general collection of relevant media bias and media bias-related tasks, more suitable for our MTL approach. We show our task selection process in Figure 3.

First, we manually assess a list of 115 media bias-related datasets in English language, categorized into task families by Wessel et al. (2023). A task family is a conceptual grouping of tasks that share similar objectives, such as those related to gender bias, encompassing pronoun coreference resolution, gender classification, and sexism detection.

We use this notion of task families to analyze general knowledge transfer between media bias tasks in Section 4.3, such as Aribandi et al. (2021b) proposed for general NLP tasks.

We filter the collection of the 115 datasets based on the following criteria[3]:

- Accessibility: Datasets have to be publicly accessible.
- Text granularity: We only use datasets labeled on a sentence level or its fragments (tokens)(not on, for instance, article level)
- Quality of annotations: We exclude datasets with no documentation, low annotation agreement or employment of machine annotation.
- Duplicates: We filter out datasets that contain full or partial duplicates of each other

Of the 115 datasets collected, we discard 11 datasets that are not publicly available. We discard 52 with article-level annotations and 5 with annotations on other levels[4]. We remove 5 datasets due to or unreliable source of annotations and discard 4 duplicates. Applying these criteria leaves 38 datasets. Including 8 handpicked datasets not

[2]BABE provides high-quality labels that capture a broad range of linguistic biases, thus allowing us to evaluate our model's generalizability within a single dataset context.

[3]We acknowledge that determining the dataset quality remains a manual and subjective choice.

[4]One discarded dataset provides only a list of biased words, other two annotations on users, and three on outlets.

originally listed gives us 46 datasets. These are categorized into task families ensuring no overlap and more than two datasets per family. Finally, ten datasets with multi-level annotations, e.g., token and sentence level, are split into tasks, yielding final number of 59 tasks.

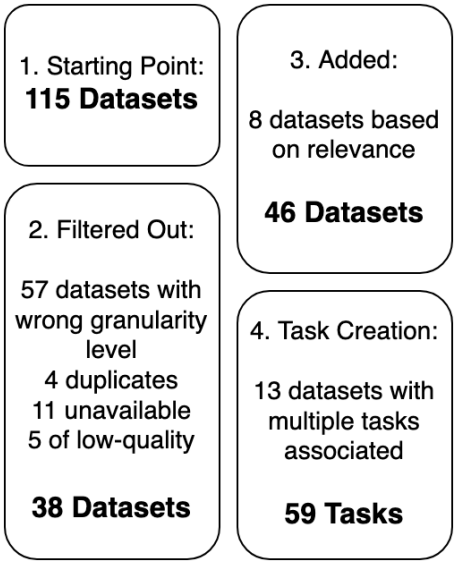


Figure 3: The workflow of collecting datasets from the initial list to the final LBM collection.

In order to standardize examples from various domains, we apply a unified text-cleaning procedure to each dataset. The process involves: (1) discarding sentences with fewer than 20 characters, (2) removing URL links, special characters, and whitespace, and (3) eliminating duplicates potentially created with steps (1) and (2).

The final Large Bias Mixture (LBM) includes 59 tasks, categorized into 9 distinct task families, encompassing 1,210,084 labeled sentences. We make the LBM publicly accessible, to facilitate research in media bias detection and other computational-social-science tasks. References and short descriptions of all datasets and corresponding tasks and task families can be found in Table 4.

3.2. The Base Model

In terms of the base language model for our procedure, we adopt a pre-trained RoBERTa (Liu et al., 2019) encoder due to its proven state-of-the-art performances across various media bias applications (Spinde et al., 2021c; Krieger et al., 2022).

3.3. The MTL framework

Pre-finetuning. To effectively harness the generalization abilities of MTL for media bias detection, we adopt a pre-finetuning procedure (Aghajanyan et al., 2021). Pre-finetuning is a complementary approach to pre-training, where a model, already pre-trained on a typically unsupervised task, is subjected to another intermediate step of pre-training. While incorporating MTL directly into the pre-training stage has demonstrated performance gains (Aribandi et al., 2021b), we opt for pre-finetuning as it offers significantly reduced computational demands while still capitalizing on the benefits of MTL (Aghajanyan et al., 2021).

Sharing representations. We use hard parameter sharing to share the underlying encoder among all tasks while using separate task-specific heads. For each task, we attach dense layers, or "heads", to the shared encoder. These heads are optimized individually per task while the shared encoder learns general bias representations.

However, multi-task optimization presents challenges due to differing gradient directions and magnitudes (Yu et al., 2020). For instance, two tasks, A and B, may have opposing gradients with the same magnitude, nullifying their sum. On the other hand, if Task A's gradient greatly surpasses that of Task B, gradient A becomes dominant. We counter the gradient misalignment by using a variation of the PCGrad de-confliction algorithm and loss scaling (Yu et al., 2020).

Conflicting gradients and loss scaling. In multi-task training involving n tasks, encoder parameters receive n potentially conflicting gradients. Efficient handling of this conflict, such as PCGrad (Yu et al., 2020), requires storing a set of gradients for each task involved in the update, leading to infeasible memory requirements among our 59 LBM tasks. Therefore, we propose a variation of PCGrad we call *PCGrad-online* which preserves the fundamental idea of the original algorithm but is more memory efficient, requiring only one set of gradients instead of n sets per update. Adopting Muppet's method, we solve the issue of varying task gradient magnitudes by re-scaling the task loss with the inverse log size of its output space, ensuring balanced gradients and preventing task dominance in training steps (Aghajanyan et al., 2021).

3.3.1. Data sampling and early stopping

To prevent large tasks from dominating the optimization, we ensure uniform data distribution by sampling one fixed-size sub-batch from each task per training step, a regular approach in MTL (Aribandi et al., 2021b; Spinde et al., 2022). We also employ early stopping as a regularization for each task individually to prevent over-fitting of tasks that converge faster. However, these methods often fall short when confronted with tasks of varied complexity and differing convergence speeds, which both is the case for tasks in LBM. When task A stops early while task B takes longer to converge, the latent

representation of the shared encoder shifts toward task B.

We aim to mitigate this issue by employing a training strategy that tackles the latent representation shift using two complementary approaches:

1. Head-Specific-Early-Stopping (HSES)

2. Resurrection

HSES. When the task stops, we stop updating its specific head parameters while still backpropagating its language model gradients. This method stems from the observation that not all tasks benefit from the shared layers' continuous learning, especially after they have reached an optimal state.

Resurrection. When the task stops, we allow it to resume the training after its validation loss starts to increase again. This enables the task to adapt its head parameters to the latest latent representation.

HSES maintains quality of faster-converging tasks, while Resurrection allows further adaptation when needed. Their combination aims for balanced, adaptive learning for tasks with varied complexities and convergence speeds. We perform a preliminary evaluation of the effectiveness of this method in the in Section 4.4. However, we stress the need for further extensive analysis in Section 5.

3.3.2. Auxiliary task selection

Multi-Task Learning often involves selecting well-used datasets, leading to potential selection biases. Furthermore, manually handpicking datasets becomes challenging due to varying and potentially ambiguous bias annotation schemes.

To automate the process of task selection, we utilize the GradTS algorithm (Ma et al., 2021). We choose GradTS due to its demonstrated efficiency and its simplicity of implementation, which enhances its usability.

Gradient-based Task selection. In line with GradTS, we construct auxiliary tasks as follows. We individually train all tasks, accumulate absolute gradients, then extract and layer-wise normalize these in the attention heads, forming a 12x12 importance matrix[5] for each task. Tasks are sorted by correlation between each task's matrix and BABE task's matrix.

We pre-finetune $m-1$ models on the first k tasks from the sorted list, where k varies from 1 to $m-1$ and m is the size of LBM. The BABE task is then evaluated on these pre-finetuned models, with the optimal k determined by evaluation loss. For further details on the GradTS algorithm, please see Ma et al. (2021).

[5] RoBERTa has 12 attention heads on each of the 12 layers.

3.3.3. Experimental setup

We split the BABE dataset into train, dev and test split with 70%,15%,15% portions respectively. For both hyperparameter tuning and auxiliary task selection, we evaluate on dev split and only use test split for the final evaluation (Section 4.2).

As fine-tuning transformers on small datasets often leads to inconsistent results, such as high performance variance (Dodge et al., 2020), we use a fixed random seed 321 for all runs in auxilliary task selection.

For the final evaluation on the test set (Table 2) we evaluate the models using 30 random seeds and report the average performance, to minimize the influence of random weight initializations. We use values $0..29$.

For optimizing the models, we use a per-task batch size of 32, an AdamW optimizer, and a polynomial learning scheduler. We run all experiments on 1 NVIDIA TESLA A100 GPU with 40 GB of memory.

4. Empirical Results

In this section, we present the results of our MTL approach. First, we report the set of auxiliary tasks selected by GradTS for pre-finetuning. Next, we assess how the model pre-finetuned on the GradTS set performs during subsequent finetuning on the BABE task, compared to a random choice of tasks and a full set of LBM tasks. We also compare the MTL approach to a single-task baseline and multiple MTL baselines and evaluate the performance of our best model, MAGPIE, on the MBIB benchmark. Then, we analyze the LBM taxonomy through a study on knowledge transfer between families. Lastly, we evaluate the effects of the proposed methods, HSES and Resurrection, through a preliminary study.

4.1. Auxiliary tasks selection

We select suitable auxiliary tasks by calculating the correlation between attention-head importance matrices. We use Kendall's correlation coefficient, as suggested by Puth et al. (2015). We find a local minimum for the BABE evaluation loss when pre-finetuning on the first $k = 10$ most correlated tasks. The final set of the ten most correlated tasks referred to as *gradts* set, is displayed in Table 1. The tasks in the *gradts* set demonstrate a strong semantic connection to media bias, encompassing areas such as lexical bias, rumor detection, and fake news detection.

Task Type	τ (correlation)
Persuasive techniques	0.73
Lexical/Informational bias	0.72
Rumour detection	0.69
Sentiment analysis	0.68
Global warming stance detection	0.68
Subjective bias	0.67
Veracity classification	0.64
Gender bias	0.64
Fake news detection	0.63

Table 1: Attention importance correlation wrt media bias task.

4.2. Evaluation

First, we finetune a model pre-finetuned on three different multi-task sets on the BABE task and compare it against the single-task RoBERTa baseline. The multi-task sets are the following:

- **MTL:Random** - Random Subset of 10 tasks
- **MTL:GradTS** - Subset of 10 tasks selected by GradTS algorithm
- **MTL:All** - Set of all tasks

We also evaluate the model pre-finetuned on the set of all tasks on MBIB. We follow the guidelines set by Wessel et al. (2023) for the evaluation. Given that MAGPIE's pre-training data includes portions of the MBIB data, we ensure that the test set for each task in MBIB is not exposed to the model during its training or validation phases.

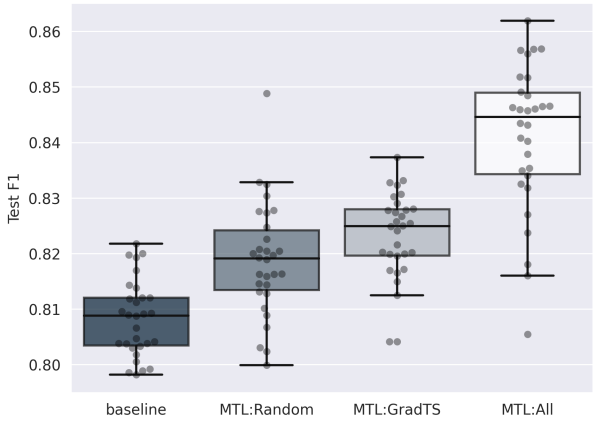


Figure 4: Final F1 score on a BABE test set averaged over 30 random seeds. All three MTL approaches outperform baseline STL finetuning. Pre-finetuning on all LBM tasks results in significantly improved performance.

Multi-Task Learning Performance. Table 2 summarizes our performance results on the BABE dataset. We observe that all of our MTL pre-finetuning schemas lead to performance improvements. In particular, pre-finetuning on **all** tasks from LBM yields a SOTA performance on the BABE dataset, achieving an 84.1% F1 score and a relative improvement of 3.3% compared to the previous baseline by Spinde et al. (2021c). While both MTL baselines - Muppet (Aghajanyan et al., 2021) and UnifiedM2 (Lee et al., 2021b) outperform single-task baseline, they underperform all of our MTL models.

On MBIB benchmark, MAGPIE ranks first on 5 out of 8 tasks. However, the improvements are only marginal. The results can be found in Appendix in Table 5.

Task scaling. GradTS task selection outperforms random tasks on average performance, yet our experiment suggests task number scaling is more crucial. This is consistent with Muppet and ExT5 results (Aghajanyan et al., 2021; Aribandi et al., 2021b), indicating MTL can compensate for scarce high-quality media bias datasets through general bias representation from other tasks. It also supports Kirstein et al. (2022)'s finding that sufficient related tasks can substitute the original task.

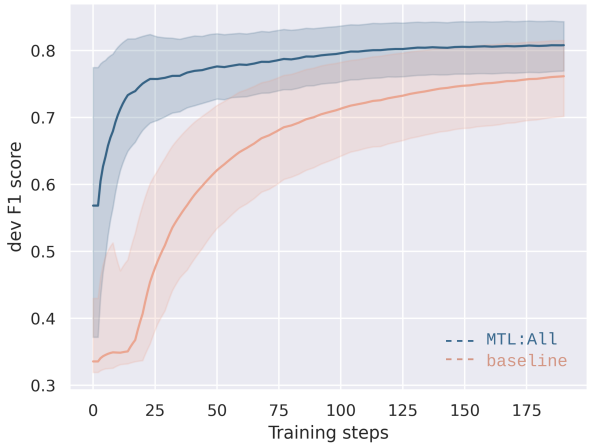


Figure 5: Evaluation F1 score during the final fine-tuning where MTL: All shows superior performance in training-step efficiency. The values are averaged over 30 random seeds. The bands mark the lowest and highest values.

Step efficiency. In addition to the performance improvements achieved, we also assess our model training efficiency. In Figure 5, we show the F1 score on the development set for the BABE task, averaged over all 30 runs. Our findings show that Multi-Task Learning only requires ~15% of the training steps used in single-task finetuning on BABE.

Model	F1	Acc	loss
Baseline (RoBERTa base)	80.83 (±0.69)	81.19 (±0.69)	43.6 (±3.54)
DA-RoBERTa	77.83 (±1.4)	78.56 (±1.3)	47.84 (±2.97)
MUPPET	80.56 (±1.3)	81.18 (±1.16)	44.19 (±4.65)
UnifiedM2	81.91 (±0.91)	82.41 (±0.88)	44.86 (±3.99)
MTL:Random	81.88 (±1.02)	82.28 (±0.97)	40.35 (±1.73)
MTL:GradTS	82.32 (±0.79)	82.64 (±0.8)	40.96 (±2.36)
MTL:All	**84.1** (±1.33)	**84.44** (±1.25)	**39.46** (±2.41)

Table 2: Performance of two MTL baseline models (Muppet, UnifiedM2) two single-task baslines (RoBERTa and DA-RoBERTa) and our three MTL models, on fine-tuning on BABE dataset and evaluating on the held-out test set. The results are averaged over 30 random seeds.

This result demonstrates the high training-step efficiency of MAGPIE in media bias classification, making MTL implementations in the media bias domain more viable in the future.

4.3. LBM taxonomy analysis

In Section 3.1, following Aribandi et al. (2021b), we introduce data task families. Aribandi et al. (2021b) uses task families for selection and knowledge transfer. To assess task families' significance in LBM taxonomy, we train each pair of families together, investigating knowledge transfer.

To account for potential negative transfer within families, we first calculate the average transfer within each family and use it as a baseline for measuring transfer between families. We train tasks from the same family together and report the average change in task performance, as depicted in Figure 6. Negative knowledge transfer is prevalent across most of our task families. However, we observe two exceptions: the hate-speech and stance detection families, where multi-task training leads to an average improvement in performance.

Next, we measure knowledge transfer between families by training each pair together. We report the average impact of each family on others and the average benefit each family gets through training with others, summarized in Section 4.3.

Our results show that, on average, the Emotionality and Sentiment analysis families provide positive transfer learning to other families. Conversely, we observe that the Fake News family benefits from knowledge transfer from every other family, with an average improvement of 1.7%. On the other hand, the Emotionality family is significantly impaired by negative transfer from other task families.

The full table of transfers can be found in Figure 7. Considering that only two families show positive transfer learning, with marginal effects of 0.11% and 0.34%, we conclude that the task families used in the construction of LBM are generally unsuitable for effectively utilizing knowledge transfer. We discuss this again in Section 5.

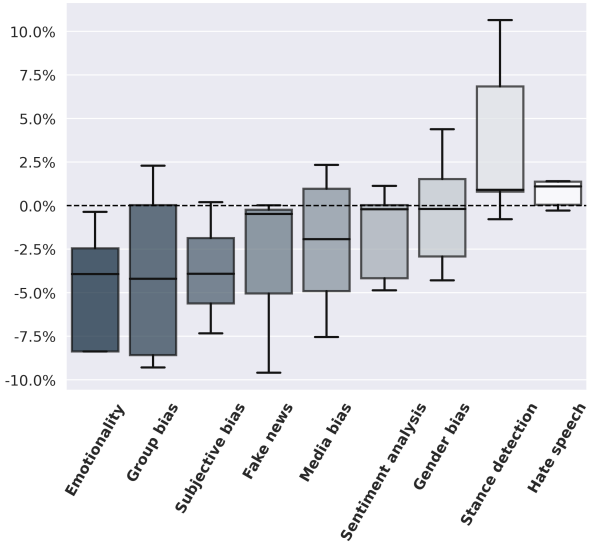


Figure 6: Average performance change per task family. Stance detection and hate speech are the only families, on average, benefitting from Multi-Task Learning.

Task Family	Transfer **from**	Transfer **to**
media bias	-2.07%	-0.94%
subjective bias	-1.26%	0.89%
hate speech	-0.87%	0.17%
gender bias	-1.01%	-1.07%
sentiment analysis	0.11%	0.72%
fake news	-0.13%	1.79%
group bias	-1.04%	0.09%
emotionality	0.34%	-6.56%
stance detection	-0.79%	-1.83%

Table 3: Evaluation of averaged transfer between task families.

4.4. Resurrection and HSES evaluation

To evaluate the Resurrection and HSES methods in combination with other training strategies, we run a grid search on the following training strategies: *HSES*, *Resurrection*, *Loss Scaling* and *Gradient Aggregation*. We calculate the average evalua-

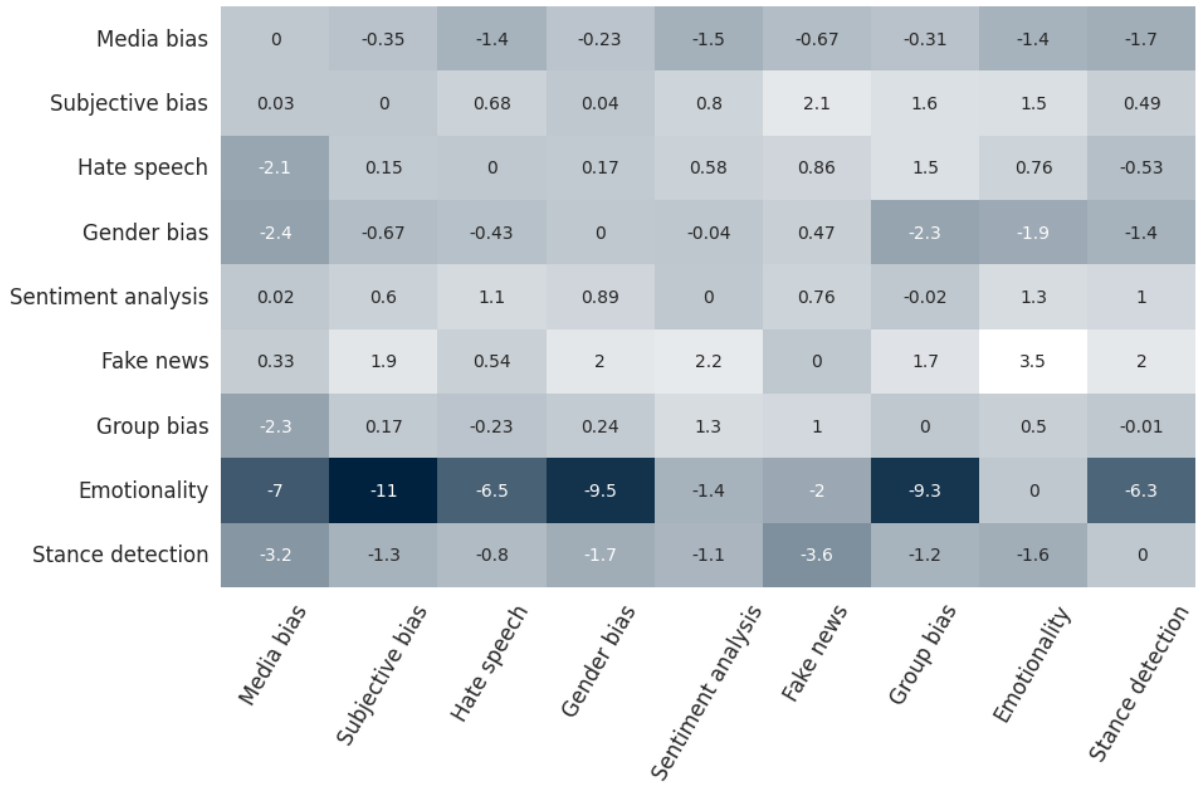
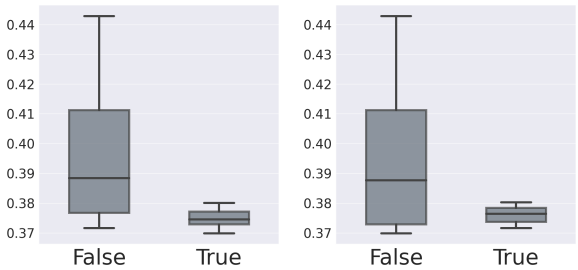


Figure 7: Average performance change when training tasks together. The change is measured with respect to transfer within each respective family (see results in section Figure 6.). The values in the horizontal axis represent the received transfer by the family on the y-axis. E.g., when training Emotionality and Subjective bias together, Emotionality gets worse by 11% whereas Subjectivity improves by 1.5%.

tion loss for both Resurrection and HSES methods across 20 tasks randomly selected from the LBM collection. The boxplot in Figure 8 shows that both methods reduce the loss by 5% and decrease the variance across different training setups by 85%. However, we hypothesize that a random constellation of tasks[6] can have a non-trivial effect on the evaluation of our technique; thus, we opt for robust examination of the methods in future work.



(a) Resurrection (b) HSES

Figure 8: An averaged evaluation loss of 20 tasks when trained with different training strategies. Both Resurrection and HSES approaches, compared to the vanilla setting, lead to significantly lower variance and overall lower loss.

5. Conclusion

This paper contributes to media bias detection by the development of MAGPIE, our large-scale multi-task learning (MTL) approach that enhances traditional models with a multi-task structure.

Additionally, we present Large Bias Mixture (LBM), a compilation of 59 bias-related tasks. This broad collection serves as a resource for the pre-training of MAGPIE. To the best of our knowledge, it is the first available MTL resource tailored specifically towards media bias.

Our study investigates the dynamics of transfer learning across tasks and task families within media bias. Despite the occurrence of negative transfer among several tasks, scaling the pre-training setup to all collected tasks in Multi-Task Learning results in a 3.3% improvement over the previous state-of-the-art, making it the biggest advancement in neural media bias classification so far. Furthermore, we report that finetuning MAGPIE on the BABE dataset only requires 15% of steps compared to RoBERTa single-task approaches. These findings underscore the effectiveness and potency of

[6] Particularly variance in task sizes and quality.

Multi-Task Learning in highly specific classification domains such as media bias.

While results suggest benefits in scaling tasks, we see more promise in novel tasks rooted in media bias, suggesting deeper exploration over simply expanding the task spectrum. Understanding families and tasks in datasets necessitates systematic analysis of label definitions, rater agreement, and inter-relatedness of dataset creation strategies. As media bias is emerging globally, incorporating multilingual models is a natural extension.

Limitations

We acknowledge the necessity for a more comprehensive analysis of the performance of the HSES and Resurrection methods to ensure robust evaluation.

Given the significant computational resources required for a single multi-task training of all tasks, we allocated these resources towards robustly evaluating the model performance rather than conducting an in-depth analysis of the optimization techniques. Consequently, the presented methods may be less reliable and have limited applicability.

While there are various models available, previous research suggests that RoBERTa exhibits strong performance on media bias tasks (Wessel et al., 2023; Spinde et al., 2021c). However, due to resource constraints, we were unable to explore models with different architectures and further refine our selection.

As the landscape of publicly available high-quality datasets for media bias is not as extensive as desired, we acknowledge our inability to capture all manifestations of media bias comprehensively. As mentioned in Section 5, a systematic and comprehensive analysis of the dataset landscape will be part of our future work.

Furthermore, although analyzing media bias on a sentence level enables a detailed examination of occurring biases, certain forms of bias, such as bias resulting from the omission of information, extend beyond the linguistic aspects of individual statements. Detecting such biases may require considering an external context.

We leave it to future work to investigate other levels of granularity, such as paragraph or article levels. In addition to these technical limitations, conducting a detailed analysis of agreement and label distribution in all utilized datasets will be necessary to make stronger claims about which datasets provide more reliable coverage of the overall concept of media bias. This is particularly important since media bias is a complex phenomenon that is not always easily identified during the annotation-gathering process.

Ethics Statement

Detecting (and as a result also highlighting) media bias instances can mitigate the negative effects of media bias, e.g., on collective decision-making (Baumer et al., 2015). However, media bias strongly depends on the context and individuals involved, making it a sensitive issue. Some bias forms depend on factors other than the content, e.g., a different text perception due to a reader's background. Therefore, a public classification of possible polarization and one-sidedness in the news must be performed transparently and reliably. The datasets used for training the classifier must be transparent and, ideally, constantly monitored by independent experts. To address possible implications of MAGPIE or related media bias applications cautiously, it is crucial to constantly control the classifications and datasets with recent findings on the perception of media bias, which is a main part of our ongoing and future work. To do so, we use resources such as, e.g., standardized questionnaires on media bias perception (Spinde et al., 2021a).

When bias detection is balanced and transparent, it can positively affect collective opinion formation and decision-making. We see no immediate negative ethical or societal impacts of our work. However, in addition to system transparency, we want to highlight that we believe it to be required for future datasets to report in greater detail about a manipulation protection strategy when developing, training, and presenting any media bias classifier. To ensure the validity of media bias detection systems, it is essential to prevent participants, especially in public studies, from manipulating algorithms by, for example, flagging neutral content as biased. Therefore, annotations should be compared among multiple users to ensure a higher level of trustworthiness. Most of the datasets available report only limitedly about such strategies. In open or crowdsourcing scenarios, it is important to collect user characteristics and deliberately include specific content that aims to provide obvious answers but may be answered differently when users follow certain patterns. This approach helps in detecting and mitigating potential biases introduced by participants, thereby maintaining the integrity and reliability of the media bias detection process. To ensure and propose stronger standards in the future, we aim to analyze all LBM datasets with regard to potential inherent bias in future work.

Acknowledgments

We are grateful for the financial support of this project provided by the Hanns-Seidel Foundation, the DAAD (German Academic Exchange Service,

program number 57515303 and 57515245), the Lower Saxony Ministry of Science and Culture, the VW Foundation and the AI Center of Czech Technical University. Furthermore, the authors would like to express their gratitude towards Jan Drchal for administrative support.

6. Bibliographical References

Armen Aghajanyan, Anchit Gupta, Akshat Shrivastava, Xilun Chen, Luke Zettlemoyer, and Sonal Gupta. 2021. Muppet: Massive multi-task representations with pre-finetuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5799–5811, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jigsaw/Conversation AI. 2019. Jigsaw unintended bias in toxicity classification.

Abeer ALDayel and Walid Magdy. 2021. Stance detection on social media: State of the art and trends. *Information Processing & Management*, 58(4):102597.

Vamsi Aribandi, Yi Tay, and Donald Metzler. 2021a. How reliable are model diagnostics? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1778–1785, Online. Association for Computational Linguistics.

Vamsi Aribandi, Yi Tay, Tal Schuster, Jinfeng Rao, Huaixiu Steven Zheng, Sanket Vaibhav Mehta, Honglei Zhuang, Vinh Q. Tran, Dara Bahri, Jianmo Ni, Jai Prakash Gupta, Kai Hui, Sebastian Ruder, and Donald Metzler. 2021b. ExT5: Towards Extreme Multi-Task Scaling for Transfer Learning. *CoRR*, abs/2111.10952. ArXiv: 2111.10952.

Sandeep Attree. 2019. Gendered Ambiguous Pronouns Shared Task: Boosting Model Confidence by Evidence Pooling.

Soumya Barikeri, Anne Lauscher, Ivan Vulić, and Goran Glavaš. 2021a. RedditBias: A Real-World Resource for Bias Evaluation and Debiasing of Conversational Language Models.

Soumya Barikeri, Anne Lauscher, Ivan Vulić, and Goran Glavaš. 2021b. RedditBias: A Real-World Resource for Bias Evaluation and Debiasing of Conversational Language Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1941–1955, Online. Association for Computational Linguistics.

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 *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1472–1482, Denver, Colorado. Association for Computational Linguistics.

Joachim Bingel and Anders Søgaard. 2017. Identifying beneficial task relations for multi-task learning in deep neural networks. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 164–169, Valencia, Spain. Association for Computational Linguistics.

Laura Ana Maria Bostan, Evgeny Kim, and Roman Klinger. 2020. GoodNewsEveryone: A Corpus of News Headlines Annotated with Emotions, Semantic Roles, and Reader Perception. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1554–1566, Marseille, France. European Language Resources Association.

BSI. 1973a. *Natural Fibre Twines*, 3rd edition. British Standards Institution, London. BS 2570.

BSI. 1973b. Natural fibre twines. BS 2570, British Standards Institution, London. 3rd. edn.

A. Castor and L. E. Pollux. 1992. The use of user modelling to guide inference and learning. *Applied Intelligence*, 2(1):37–53.

Shijie Chen, Yu Zhang, and Qiang Yang. 2021. Multi-Task Learning in Natural Language Processing: An Overview. *arXiv:2109.09138 [cs]*.

J.L. Chercheur. 1994. *Case-Based Reasoning*, 2nd edition. Morgan Kaufman Publishers, San Mateo, CA.

N. Chomsky. 1973. Conditions on transformations. In *A festschrift for Morris Halle*, New York. Holt, Rinehart & Winston.

Costanza Conforti, Jakob Berndt, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier. 2020. Will-they-won't-they: A very large dataset for stance detection on Twitter. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1724, Online. Association for Computational Linguistics.

Jamell Dacon and Haochen Liu. 2021. Does Gender Matter in the News? Detecting and Examining Gender Bias in News Articles. In *Companion Proceedings of the Web Conference 2021*, WWW '21, pages 385–392, New York, NY, USA. Association for Computing Machinery.

Kareem Darwish, Walid Magdy, and Tahar Zanouda. 2017. Trump vs. Hillary: What Went Viral During the 2016 US Presidential Election. In *Social Informatics*, Lecture Notes in Computer Science, pages 143–161, Cham. Springer International Publishing.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11. Issue: 1.

Emily Dinan, Angela Fan, Ledell Wu, Jason Weston, Douwe Kiela, and Adina Williams. 2020. Multi-dimensional gender bias classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 314–331, Online. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of Wikipedia: Knowledge-Powered Conversational agents.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of Wikipedia: Knowledge-powered Conversational Agents. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and Mitigating Unintended Bias in Text Classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, pages 67–73, New York, NY, USA. Association for Computing Machinery.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. Fine-Tuning Pretrained Language Models: Weight Initializations, Data Orders, and Early Stopping.

Umberto Eco. 1990. *The Limits of Interpretation*. Indian University Press.

Lisa Fan, Marshall White, Eva Sharma, Ruisi Su, Prafulla Kumar Choubey, Ruihong Huang, and Lu Wang. 2019. 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)*, pages 6343–6349, Hong Kong, China. Association for Computational Linguistics.

W. Ferreira and A. Vlachos. 2016. Emergent: a novel data-set for stance classification.

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1). Number: 1.

Michael Färber, Victoria Burkard, Adam Jatowt, and Sora Lim. 2020. A Multidimensional Dataset Based on Crowdsourcing for Analyzing and Detecting News Bias. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, pages 3007–3014, New York, NY, USA. Association for Computing Machinery. Event-place: Virtual Event, Ireland.

Akash Gautam, Puneet Mathur, Rakesh Gosangi, Debanjan Mahata, Ramit Sawhney, and Rajiv Ratn Shah. 2020. #MeTooMA: Multi-Aspect Annotations of Tweets Related to the MeToo Movement. *Proceedings of the International AAAI Conference on Web and Social Media*, 14:209–216.

Jennifer Golbeck, Zahra Ashktorab, Rashad O. Banjo, Alexandra Berlinger, Siddharth Bhagwan, Cody Buntain, Paul Cheakalos, Alicia A. Geller, Quint Gergory, Rajesh Kumar Gnanasekaran, Raja Rajan Gunasekaran, Kelly M. Hoffman, Jenny Hottle, Vichita Jienjitlert, Shivika Khare, Ryan Lau, Marianna J. Martindale, Shalmali Naik, Heather L. Nixon, Piyush Ramachandran, Kristine M. Rogers, Lisa Rogers, Meghna Sardana Sarin, Gaurav Shahane, Jayanee Thanki, Priyanka Vengataraman, Zijian Wan, and Derek Michael Wu. 2017. A Large Labeled Corpus for Online Harassment Research. In *Proceedings of the 2017 ACM on Web Science Conference*, pages 229–233, Troy New York USA. ACM.

Dylan Grosz and Patricia Conde-Cespedes. 2020. Automatic Detection of Sexist Statements Commonly Used at the Workplace. In Wei Lu and Kenny Q. Zhu, editors, *Trends and Applications in Knowledge Discovery and Data Mining*, volume

12237, pages 104–115. Springer International Publishing, Cham.

Felix Hamborg, Karsten Donnay, and Bela Gipp. 2019. Automated identification of media bias in news articles: an interdisciplinary literature review. *International Journal on Digital Libraries*, 20(4):391–415.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2019. An interactive multi-task learning network for end-to-end aspect-based sentiment analysis. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 504–515, Florence, Italy. Association for Computational Linguistics.

Zihao He, Negar Mokhberian, and Kristina Lerman. 2022. Infusing Knowledge from Wikipedia to Enhance Stance Detection.

Paul Gerhard Hoel. 1971a. *Elementary Statistics*, 3rd edition. Wiley series in probability and mathematical statistics. Wiley, New York, Chichester. ISBN 0 471 40300.

Paul Gerhard Hoel. 1971b. *Elementary Statistics*, 3rd edition, Wiley series in probability and mathematical statistics, pages 19–33. Wiley, New York, Chichester. ISBN 0 471 40300.

Christoph Hube and Besnik Fetahu. 2018. Detecting Biased Statements in Wikipedia. In *Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW '18*, pages 1779–1786, Lyon, France. ACM Press.

Christoph Hube and Besnik Fetahu. 2019. Neural Based Statement Classification for Biased Language. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, WSDM '19, pages 195–203, New York, NY, USA. Association for Computing Machinery. Event-place: Melbourne VIC, Australia.

Pere-Lluís Huguet Cabot, David Abadi, Agneta Fischer, and Ekaterina Shutova. 2021. Us vs. Them: A Dataset of Populist Attitudes, News Bias and Emotions. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1921–1945. Association for Computational Linguistics. Event-place: Online.

Otto Jespersen. 1922. *Language: Its Nature, Development, and Origin*. Allen and Unwin.

Frederic Thomas Kirstein, Jan Philip Wahle, Terry Ruas, and Bela Gipp. 2022. Analyzing multi-task learning for abstractive text summarization. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 54–77, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. 2018. All-in-one: Multi-task learning for rumour verification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3402–3413, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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 *2022 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, Cologne, Germany.

Maria Krommyda, Anastasios Rigos, Kostas Bouklas, and Angelos Amditis. 2021a. An Experimental Analysis of Data Annotation Methodologies for Emotion Detection in Short Text Posted on Social Media. *Informatics*, 8(1):19.

Maria Krommyda, Anastasios Rigos, Kostas Bouklas, and Angelos Amditis. 2021b. An Experimental Analysis of Data Annotation Methodologies for Emotion Detection in Short Text Posted on Social Media. *Informatics*, 8(1):19.

Nayeon Lee, Yejin Bang, Andrea Madotto, and Pascale Fung. 2021a. Mitigating Media Bias through Neutral Article Generation. *CoRR*, abs/2104.00336. _eprint: 2104.00336.

Nayeon Lee, Belinda Z. Li, Sinong Wang, Pascale Fung, Hao Ma, Wen-tau Yih, and Madian Khabsa. 2021b. On unifying misinformation detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5479–5485, Online. Association for Computational Linguistics.

Sora Lim, Adam Jatowt, Michael Färber, and Masatoshi Yoshikawa. 2020. Annotating and Analyzing Biased Sentences in News Articles using Crowdsourcing. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1478–1484, Marseille, France. European Language Resources Association.

Bing Liu. 2012. *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies. Springer International Publishing, Cham.

Rui Liu, Zheng Lin, Yutong Tan, and Weiping Wang. 2021. Enhancing Zero-shot and Few-shot Stance Detection with Commonsense Knowledge Graph. In *Findings of the Association for Computational*

Linguistics: ACL-IJCNLP 2021, pages 3152–3157, Online. Association for Computational Linguistics.

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. *CoRR*, abs/1907.11692. Eprint: 1907.11692.

Yiwei Luo, Dallas Card, and Dan Jurafsky. 2020. DeSMOG: Detecting Stance in Media On Global Warming. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3296–3315, Online. Association for Computational Linguistics.

Weicheng Ma, Renze Lou, Kai Zhang, Lili Wang, and Soroush Vosoughi. 2021. GradTS: A gradient-based automatic auxiliary task selection method based on transformer networks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5621–5632, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011a. Learning Word Vectors for Sentiment Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011b. Learning Word Vectors for Sentiment Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14867–14875.

Alexander Miller, Will Feng, Dhruv Batra, Antoine Bordes, Adam Fisch, Jiasen Lu, Devi Parikh, and Jason Weston. 2017. ParlAI: A dialog research software platform. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 79–84, Copenhagen, Denmark. Association for Computational Linguistics.

Nailia Mirzakhmedova, Johannes Kiesel, Milad Alshomary, Maximilian Heinrich, Nicolas Handke, Xiaoni Cai, Barriere Valentin, Doratossadat Dastgheib, Omid Ghahroodi, Mohammad Ali Sadraei, Ehsaneddin Asgari, Lea Kawaletz, Henning Wachsmuth, and Benno Stein. 2023. The touché23-valueeval dataset for identifying human values behind arguments.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41.

Saif M. Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2017. Stance and Sentiment in Tweets. *ACM Transactions on Internet Technology*, 17(3):26:1–26:23.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021a. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021b. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371. Association for Computational Linguistics. Event-place: Online.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Bo Pang and Lillian Lee. 2004. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04, pages 271–es, USA. Association for Computational Linguistics. Event-place: Barcelona, Spain.

Jakub Piskorski, Nicolas Stefanovitch, Valerie-Anne Bausier, Nicolo Faggiani, Jens Linge,

Sopho Kharazi, Nikolaos Nikolaidis, Giulia Teodori, Bertrand De Longueville, Brian Doherty, Jason Gonin, Camelia Ignat, Bonka Kotseva, Eleonora Mantica, Lorena Marcaletti, Enrico Rossi, Alessio Spadaro, Marco Verile, Giovanni Da San Martino, Firoj Alam, and Preslav Nakov. 2023. News categorization, framing and persuasion techniques: Annotation guidelines. Technical report, European Commission Joint Research Centre, Ispra (Italy).

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 Task 4: Aspect Based Sentiment Analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

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*, volume 34, pages 480–489. Issue: 01.

Rajkumar Pujari, Erik Oveson, Priyanka Kulkarni, and Elnaz Nouri. 2022. Reinforcement Guided Multi-Task Learning Framework for Low-Resource Stereotype Detection. Technical report. ADS Bibcode: 2022arXiv220314349P Type: article.

Marie-Therese Puth, Markus Neuhäuser, and Graeme D. Ruxton. 2015. Effective use of Spearman's and Kendall's correlation coefficients for association between two measured traits. *Animal Behaviour*, 102:77–84.

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. *Journal of Machine Learning Research*, 21(140):1–67.

Shaina Raza, Deepak John Reji, and Chen Ding. 2022. Dbias: Detecting biases and ensuring fairness in news articles. *International Journal of Data Science and Analytics*.

Carley Reardon, Sejin Paik, Ge Gao, Meet Parekh, Yanling Zhao, Lei Guo, Margrit Betke, and Derry Wijaya. 2022. BU-NEmo: an Affective Dataset of Gun Violence News. *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, pages 2507–2516.

Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. 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)*, volume 1, pages 1650–1659, Sofia, Bulgaria. Association for Computational Linguistics.

Terry Ruas, Charles P. H. Ferreira, William Gorsky, Fabrício O. França, and Débora M. R. Medeiros. a. Enhanced word embeddings using multi-semantic representation through lexical chains. 532:16–32.

Terry Ruas, William Gorsky, and Akiko Aizawa. b. Multi-sense embeddings through a word sense disambiguation process. 136:288–303.

Sebastian Ruder and Barbara Plank. 2017. Learning to select data for transfer learning with Bayesian optimization. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 372–382, Copenhagen, Denmark. Association for Computational Linguistics.

Niloofar Safi Samghabadi, Parth Patwa, Srinivas PYKL, Prerana Mukherjee, Amitava Das, and Thamar Solorio. 2020. Aggression and Misogyny Detection using BERT: A Multi-Task Approach. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 126–131, Marseille, France. European Language Resources Association (ELRA).

Mattia Samory, Indira Sen, Julian Kohne, Fabian Floeck, and Claudia Wagner. 2020. "Call me sexist, but...": Revisiting Sexism Detection Using Psychological Scales and Adversarial Samples.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020a. FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media. *Big Data*, 8(3):171–188.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020b. FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media. *Big Data*, 8(3):171–188.

Charles Joseph Singer, E. J. Holmyard, and A. R. Hall, editors. 1954–58. *A history of technology*. Oxford University Press, London. 5 vol.

Manjira Sinha and Tirthankar Dasgupta. 2021. Determining Subjective Bias in Text through Linguistically Informed Transformer based Multi-Task Network. In *Proceedings of the 30th ACM International Conference on Information & Knowledge*

Management, pages 3418–3422. ACM. Event-place: Virtual Event Queensland Australia.

Parinaz Sobhani, Diana Inkpen, and Xiaodan Zhu. 2017. A Dataset for Multi-Target Stance Detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 551–557, Valencia, Spain. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013a. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013b. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Timo Spinde, Smilla Hinterreiter, Fabian Haak, Terry Ruas, Helge Giese, Norman Meuschke, and Bela Gipp. 2023. The media bias taxonomy: A systematic literature review on the forms and automated detection of media bias. *arXiv preprint arXiv:2312.16148*.

Timo Spinde, Christina Kreuter, Wolfgang Gaissmaier, Felix Hamborg, Bela Gipp, and Helge Giese. 2021a. 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)*, pages 61–69.

Timo Spinde, David Krieger, Manu Plank, and Bela Gipp. 2021b. Towards A Reliable Ground-Truth For Biased Language Detection. In *Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)*, Virtual Event.

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*, Virtual event. Tex.pubstate: published tex.tppubtype: inproceedings.

Timo Spinde, Manuel Plank, Jan-David Krieger, Terry Ruas, Bela Gipp, and Akiko Aizawa. 2021c. Neural Media Bias Detection Using Distant Supervision With BABE - Bias Annotations By Experts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1166–1177, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Dhanya Sridhar and Lise Getoor. 2019. Estimating causal effects of tone in online debates. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, IJCAI'19, page 1872–1878. AAAI Press.

Jannik Strötgen and Michael Gertz. 2012. Temporal tagging on different domains: Challenges, strategies, and gold standards. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3746–3753, Istanbul, Turkey. European Language Resource Association (ELRA).

S. Superman, B. Batman, C. Catwoman, and S. Spiderman. 2000. *Superheroes experiences with books*, 20th edition. The Phantom Editors Associates, Gotham City.

Rob Voigt, David Jurgens, Vinodkumar Prabhakaran, Dan Jurafsky, and Yulia Tsvetkov. 2018a. RtGender: A Corpus for Studying Differential Responses to Gender. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Rob Voigt, David Jurgens, Vinodkumar Prabhakaran, Dan Jurafsky, and Yulia Tsvetkov. 2018b. RtGender: A Corpus for Studying Differential Responses to Gender. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jan Wahle, Bela Gipp, and Terry Ruas. a. Paraphrase Types for Generation and Detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12148–12164. Association for Computational Linguistics.

Jan Philip Wahle, Terry Ruas, Frederic Kirstein, and Bela Gipp. b. How Large Language Models are

Transforming Machine-Paraphrase Plagiarism. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 952–963. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3261–3275.

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*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

William Yang Wang. 2017a. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

William Yang Wang. 2017b. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the GAP: A Balanced Corpus of Gendered Ambiguous Pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617.

Maxwell Weinzierl and Sanda Harabagiu. 2022. VaccineLies: A Natural Language Resource for Learning to Recognize Misinformation about the COVID-19 and HPV Vaccines. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6967–6975, Marseille, France. European Language Resources Association.

Martin Wessel, Tomas Horych, Terry Ruas, Akiko Aizawa, Bela Gipp, and Timo Spinde. 2023. Introducing MBIB - The First Media Bias Identification Benchmark Task and Dataset Collection. In *Proceedings of 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'23)*, New York, NY, USA. ACM. ISBN 978-1-4503-9408-6/23/07.

Theresa Ann Wilson. 2008. *Fine-Grained Subjectivity and Sentiment Analysis: Recognizing the Intensity, Polarity, and Attitudes of Private States*. Ph.D. thesis, USA. AAI3322382.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017a. Ex Machina: Personal Attacks Seen at Scale. In *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017*, pages 1391–1399. ACM.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017b. Ex Machina: Personal Attacks Seen at Scale. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, pages 1391–1399, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. 2020. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 5824–5836. Curran Associates, Inc.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015a. Character-level convolutional networks for text classification. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pages 649–657, Cambridge, MA, USA. MIT Press.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015b. Character-level Convolutional Networks for Text Classification. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 649–657.

A. Appendix

Task Family	Dataset	# sentences	Task
Subjective bias	SUBJ (Pang and Lee, 2004)	10.000	Binary Classification
	Wiki Neutrality Corpus (Pryzant et al., 2020)	52.036	Token-level Classification
	NewsWCL50 (Hamborg et al., 2019)	731	Regression
	CW_HARD (Hube and Fetahu, 2019)	6.843	Binary Classification
News bias	MultiDimNews (Färber et al., 2020)	2.015	Multi-Label Classification
	BASIL (Fan et al., 2019)	7.987	Multi-Class Classification
			Token-Level Classification
	Starbucks (Lim et al., 2020)	866	Regression
	SemEval2023Task3 (Piskorski et al., 2023)	5.219	Binary Classification
	BABE (Spinde et al., 2021c)	3.672	Binary Classification
			Token-Level Classification
Hate speech	OffensiveLanguage (Davidson et al., 2017)	23.198	Multi-Class Classification
	OnlineHarassmentDataset (Golbeck et al., 2017)	19.613	Binary Classification
	WikiDetoxToxicity (Wulczyn et al., 2017a)	138.827	Regression
	WikiDetoxAggression (Wulczyn et al., 2017a)	101.159	Binary Classification
			Regression
	Jigsaw (AI, 2019)	101.060	Binary Classification
	MeTooMA (Gautam et al., 2020)	7.388	Multi-Label Classification
	WikiMadlibs (Dixon et al., 2018)	74.972	Binary Classification
	HateXplain (Mathew et al., 2021)	18.962	Multi-Class Classification
			Multi-Label Classification
			Token-Level Classification
	HateSpeechTwitter (Founta et al., 2018)	48.572	Multi-Class Classification
Gender bias	GAP (Webster et al., 2018)	4.373	Multi-Class Classification
	RtGender (Voigt et al., 2018a)	21.690	Binary Classification
	MDGender (Dinan et al., 2020)	2.332	Multi-Class Classification
	TRAC2 (Safi Samghabadi et al., 2020)	3.983	Binary Classification
	Funpedia (Miller et al., 2017)	11.256	Multi-Class Classification
	WizardsOfWikipedia (Dinan et al., 2019)	29.777	Multi-Class Classification
Sentiment analysis	SST2 (Socher et al., 2013a)	9.436	Binary Classification
	IMDB (Maas et al., 2011a)	13.139	Binary Classification
	MPQA (Wilson, 2008)	3.508	Binary Classification
	SemEval2014 (Pontiki et al., 2014)	5.794	Token-Level Classification
	AmazonReviews (Zhang et al., 2015b)	167.396	Binary Classification
Fake news	LIAR (Wang, 2017a)	12.742	Regression
	FakeNewsNet (Shu et al., 2020a)	21.299	Binary Classification
	PHEME (Kochkina et al., 2018)	5.022	Binary Classification
			Multi-Class Classification
Emotionality	GoodNewsEveryone (Bostan et al., 2020)	4.428	Token-Level Classification
			Token-Level Classification
	BU-NEMO (Reardon et al., 2022)	12.576	Multi-Class Classification
	EmotionTweets (Krommyda et al., 2021b)	195.744	Multi-Class Classification
	DebateEffects (Sridhar and Getoor, 2019)	6.941	Regression
Group bias	CrowSPairs (Nangia et al., 2020)	3.009	Binary Classification
			Multi-Class Classification
			Token-Level Classification
	StereoSet (Nadeem et al., 2021b)	4.170	Binary Classification
			Multi-Class Classification
	StereotypeDataset (Pujari et al., 2022)	2.208	Binary Classification
			Multi-Label Classification
	RedditBias (Barikeri et al., 2021b)	10.395	Binary Classification
			Multi-Class Classification
			Token-Level Classification
Stance detection	SemEval2023Task4 (Mirzakhmedova et al., 2023)	5.219	Binary Classification
	VaccineLies (Weinzierl and Harabagiu, 2022)	4.497	Multi-Class Classification
	SemEval2016Task6 (Mohammad et al., 2016)	4.849	Multi-Class Classification
	WTWT (Conforti et al., 2020)	24.681	Multi-Class Classification
	MultiTargetStance (Sobhani et al., 2017)	4.430	Multi-Class Classification
	GWSD (Luo et al., 2020)	2.010	Multi-Class Classification
		$\sum$ 1.210.084	

Table 4: References and description to all 59 Tasks (46 datasets) in LBM collection.

MBIB Task	MAGPIE	RoBERTa	ConvBERT
Linguistic Bias	**0.7139**	0.7076	0.7126
Cognitive Bias	**0.7086**	0.7037	0.7044
Text-Level Context Bias	0.7638	0.7646	0.7697
Hatespeech	0.8747	0.8759	0.8805
Gender Bias	**0.8344**	0.8322	0.8257
Racial Bias	**0.8809**	0.8761	0.8772
Fake News	0.6709	0.6711	0.6787
Political Bias	**0.7059**	0.7029	0.7041

Table 5: Performance of MAGPIE and two baselines RoBERTa, and ConvBERT (state-of-the-art model in Wessel et al. (2023)) on the Media Bias Identification Benchmark (MBIB) tasks.