Introducing MBIB - the first Media Bias Identification Benchmark Task and Dataset Collection

Martin Wessel
University of Konstanz
Konstanz, Germany
m.wessel@media-bias-research.org

Tomáš Horych
Czech Technical University
Prague, Czech Republic
t.horych@media-bias-research.org

Terry Ruas
University of Göttingen
Göttingen, Germany
ruas@uni-goettingen.de

Akiko Aizawa
National Institute of Informatics
Tokyo, Japan
aizawa@nii.ac.jp

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

Timo Spinde
University of Göttingen
Göttingen, Germany
t.spinde@media-bias-research.org

arXiv:2304.13148v1 [cs.IR] 25 Apr 2023

ABSTRACT

Although media bias detection is a complex multi-task problem, there is, to date, no unified benchmark grouping these evaluation tasks. We introduce the Media Bias Identification Benchmark (MBIB), a comprehensive benchmark that groups different types of media bias (e.g., linguistic, cognitive, political) under a common framework to test how prospective detection techniques generalize. After reviewing 115 datasets, we select nine tasks and carefully propose 22 associated datasets for evaluating media bias detection techniques. We evaluate MBIB using state-of-the-art Transformer techniques (e.g., T5, BART). Our results suggest that while hate speech, racial bias, and gender bias are easier to detect, models struggle to handle certain bias types, e.g., cognitive and political bias. However, our results show that no single technique can outperform all the others significantly. We also find an uneven distribution of research interest and resource allocation to the individual tasks in media bias. A unified benchmark encourages the development of more robust systems and shifts the current paradigm in media bias detection evaluation towards solutions that tackle not one but multiple media bias types simultaneously.

CCS CONCEPTS

• **Computing methodologies** → **Information extraction**; *Natural language processing*; • **Information systems** → **Content analysis and feature selection**; *Language models.*

KEYWORDS

media bias detection, benchmark, model comparison, datasets

1 INTRODUCTION

Media bias is often related to content favoring a particular viewpoint or ideology (e.g., political) [34]. Such bias has been the focus of various research projects [45] and is generally defined and summarized by the term **media bias** [44]. Media bias can have various negative impacts, e.g., the distribution of false facts, affecting decision-making processes, and endangering the readers' trust in news [12]. Accordingly, drawing attention to instances of media bias can have far-reaching benefits and counterbalance its impact [4]. Although it may not be possible to eliminate media bias completely, acknowledging its presence helps educate readers about it and encourages journalists and publishers to evaluate their work

impartially. Given the vast amount of digital information available, getting an overview of bias in various media outlets is only possible with automated solutions [40]. Therefore, automatic media bias detection experiences steadily growing research attention.

While substantial research exists for solving specific media bias detection tasks, e.g., gender bias in the news [52], they usually concentrate on identifying a single aspect within the media bias spectrum. To the best of our knowledge, no single uniform benchmark allows for the comparison of prominent models, and no overview of potential media bias tasks and subtasks exists. The lack of a benchmark leads to some problems: (a) solutions are focused on specific subtasks and bias types, (b) no standardized comparison between models is possible, and (c) models in the domain are less likely to make use of multi-task learning, while media bias itself is a multi-task problem [47]. A unified benchmark covering different bias aspects will, in the future, also allow the development of more robust systems. Therefore, we propose the Media Bias Identification Benchmark (MBIB) to introduce a challenging media bias task collection with associated datasets. Our benchmark is composed of nine tasks and 22 associated datasets. As MBIB covers a wide range of media bias types (e.g., framing, political)[1], its solution provides a reliable proxy for evaluation covering the domain as entirely and socially relevant as possible. To curate the 22 associated datasets, we conduct an extensive literature search, concluding with a list of 115 related datasets. In the following, we briefly summarize our contributions.

(1) The first media bias benchmark is composed of nine tasks and 22 datasets.
(2) A framework for evaluating models in a standardized way.
(3) MBIB is publicly available at

https://github.com/Media-Bias-Group/MBIB[2]

2 RELATED WORK

2.1 Media bias

While media bias is generally often referred to as communicating about something "in a prejudiced manner or with slanted viewpoint" [24], many definitions for media bias and its subtypes exist

---

[1]Full list in Table 2.

[2]To facilitate easy access, we also share our MBIB base corpus on huggingface so that it can be fetched directly. We provide detailed information on the usage of all our code and data within the GitHub repository.

in the literature. For instance, definitions of hate speech in the literature differ on whether and how strongly offensive language can constitute hate speech [28]. To name another example, within linguistic bias, some works rely stronger on traditional linguistic features, focusing on bias rather as an objective entity [34]. Others use linguistic bias, or bias by word choice, as a majorly subjective concept [48], or as a general word choice communicating stereotypes [5]. In their literature review, Spinde et al. [43] show how the various definitions diverge in detail. They also introduce the media bias framework [43], a coherent overview of the current state of research on media bias from different perspectives, such as linguistic bias, text-level context bias, cognitive bias, group bias, and others. The authors show that media bias detection is a highly active research field and that transformer-based approaches have significantly improved the media bias classification task over recent years. However, most advances in media bias detection are still tied to single tasks (e.g., linguistic bias) and do not report on the generalizability of their results [9, 19, 26, 27, 48, 57]. We give a full overview of existing datasets in Section 3. Overlapping and unclear definitions of media bias and its subtypes can lead authors to explore different biases under the same name. Current system evaluations are limited without a good overview of available definitions and datasets, and results are difficult to compare [43].[3] The lack of a standardized evaluation procedure makes comparing and reproducing results even more challenging. Moreover, focusing on individual tasks does not allow for a holistic assessment of systems for media bias. Developing systems that detect media bias as a whole will increase efficiency (as it reduces the need for multiple methods), allow for consistency (as it reduces the variation between evaluations), and lead to more comprehensive bias assessments. By introducing MBIB, we aim to make such a development possible.

2.2 Language Processing Benchmarks

In other areas of natural language processing (outside of media bias), task benchmarks have shown how important proxy tasks can be when tackling complex problems. For example, the General Language Understanding Evaluation (GLUE) benchmark [56] decomposes natural language understanding into smaller tasks, formalizing the need for models with strong generalizability in the domain. GLUE introduces various language understanding tasks, such as similarity, inference, and question-answering tasks. It also provides a standardized score that can be used to compare the performance of different NLP systems across different tasks. An improved version of GLUE is SuperGLUE [55], which adds a broader variety of tasks[4] and also includes fewer training examples, making it more challenging to learn and generalize to new examples. The question answering included in SuperGLUE is increased in difficulty by requiring more contextual understanding and the combination of information from different parts of the text. For both, SuperGLUE and GLUE, a task consists of a single labeled dataset. Another benchmark follows a similar strategy for natural language

processing systems, BIG-Bench [50]. Opposed to those in existing benchmarks, media bias datasets are too small or only cover a particular aspect of a task. For example, they describe bias by word choice without a perspective on linguistic bias in total [48].

As existing benchmarks have successfully formalized a challenging problem by decomposing it into several defined subtasks, they can serve as inspiration for media bias detection. Still, to overcome the abovementioned drawbacks, we need to adapt the strategies from existing benchmarks, which we describe in Section 3.3. We base our MBIB guidelines on the best-practice benchmark examples, mainly on SuperGLUE [55]: Every task should have a well-defined metric, and all tasks should have public training data available. Even more, the tasks should be too difficult to solve for current solutions but should be solvable by humans.[5] Finally, tasks should be in a format that is as simple as possible.

3 BUILDING MBIB

In the following, we detail the process that leads to the creation of MBIB. First, we collect and select relevant media bias tasks and related datasets (Section 3.1). The tasks should cover media bias as comprehensively as possible while reflecting societal relevance and existing research priorities. For instance, gender bias in the news has become a focus of bias research [52]. Second, we refine the initial list of datasets through a detailed study (Section 3.2). Mainly, we incorporate factors such as dataset availability, size, and quality. The avoidance of duplicates (using the same data basis or the same data set) is also taken into account here. Third, we preprocess and unify the chosen datasets (Section 3.4). We also give a detailed overview of the properties of the selected datasets (Section 4). Finally, we conclude the construction of MBIB with a framework defining how models can be evaluated based on the tasks (Section 5). We set a transformer-based baseline performance within the framework on all MBIB tasks.

3.1 Media Bias Tasks

A comprehensive curation of tasks that address media bias is challenging as it encompasses various forms of bias. To ensure a systematic examination, each task should tackle an independent subcategory of media bias detection (e.g., linguistic or gender bias). Specifying media bias types into categories facilitates each task's definition, delimitation, and interpretation. We select tasks based on two criteria:

(1) When occurring in media coverage, the task's associated bias should constitute a form of media bias.
(2) Tasks should either (A) be part of the media bias framework presented by Spinde et al. [43], or (B) be an independent, distinguishable research field of societal importance.[6]

To identify task candidates, we assess a list of 322 media bias-related publications mentioned in [43].[7] To the best of our knowledge, it is the most extensive existing literature collection on media bias to date [43]. Their list categorizes the publications by the type

---

[3]The low comparability is even more surprising given that many existing works highlight how important task awareness is in resource-scarce domains [2] and in particular for media bias research [21, 46, 48].
[4]While the tasks in GLUE mainly involve sentence-level and word-level understanding, SuperGLUE tasks involve understanding and reasoning about more complex phenomena, such as temporal relations, coreference, and commonsense knowledge.

[5]This can sometimes also mean that only trained humans can solve the task.
[6]This criterion implies that the tasks are not necessarily mutually exclusive.
[7]The authors automatically and then manually filter all relevant media bias research published between 2018 and 2023 [43]. In total, they consider over 100,000 publications.

of bias they focus on. We use these (task) categories as a starting point for our task selection, given that they all fulfill criterion (1). Also, all tasks fulfill either criterion (2A) or (2B). In total, we identify the following task candidates: linguistic, text-level context, reporting-level context, cognitive bias (by criterion (2A)), as well as hate speech, fake news, political bias, racial bias, gender bias, religious bias, group bias, and framing effects (by criterion (2B)).

Of the task candidates mentioned above, framing effects, group bias, and religious bias are not included in the MBIB task. Group bias is an umbrella term describing bias introduced when dealing with various groups and includes gender bias, racial bias, and religious bias. Instead of group bias, we, therefore, include the single tasks to avoid repetitions. Framing effects refer to how media organizations present information to influence the reader's perception. We do not include framing effects due to their similarity to framing bias being a component of the linguistic bias task. We decided that religious bias is not part of MBIB due to the low research interest (only one associated paper). Also, we do not include a wide range of media bias-related tasks, such as sentiment analysis or stance detection [47] since they do not directly identify forms of media bias. Especially sentiment detection is already covered in multiple existing benchmarks [32]. The remaining five task candidates fulfilling criterion (2B) are added as external tasks to MBIB, together with the four tasks fulfilling criterion (2A). Therefore, MBIB consists of 9 tasks in total.

In the remainder of this section, we briefly introduce the chosen tasks. More details and respective subcategories of each task can be found in [43]. For every task, Table 1 shows an example.

Linguistic bias encompasses all forms of bias induced by lexical features, such as word choice and sentence structure, often subconsciously used [43]. Generally, linguistic bias is expressed through specific word choice that reflects the social-category cognition applied to any described group or individual(s) [5].

Text-level context bias refers to the expression of a text's context, whereby words and statements can shape the context of an article and sway the reader's perspective [17]. These biases can be used to portray a particular opinion in a biased way by criticizing one side more than the other, using inflammatory words, or omitting relevant information.

Reporting-level context bias refers to bias that arises through decisions made by editors and journalists on what events to report and which sources to use [7]. While text-level context bias examines the bias present within an individual article, reporting-level bias focuses on systematic attention given to specific topics.

Cognitive bias occurs when readers introduce bias by selecting which articles to read and which sources to trust, which can be amplified in social media [31]. These biases can lead to self-reinforcing cycles and expose readers to only one side of an issue.

Hate speech refers to any language that manifests hatred towards a specific group or aims to degrade, humiliate, or offend [28]. Usually, hate speech is induced by using linguistic bias [30]. Particularly in social media, the impact of hate speech is significant and exacerbates tensions between involved parties. However, similar processes can also be observed within, e.g., comments on news websites [59].

Fake news refers to published content based on false claims and premises, presented as being true to deceive the reader [51]. Research on fake news detection typically focuses on detecting it through linguistic features or comparing content to verified information [51]. Fake news have serious consequences, such as potential influences on the readers' health and political decisions [36]. However, the exact overlap between media bias and fake news is yet unclear; we address this again in Section 6.

Racial bias is expressed through negative or positive portrayals of racial groups. Research has shown that racial bias in news coverage can severely impact affected minorities, such as strengthening stereotypes and discrimination [8, 29].

Gender bias in media can manifest as discrimination against one gender through underrepresentation or negative portrayal. Gender bias in media can severely impact perceptions of professions and role models [39], as well as voting decisions [23].

Political bias refers to a text's political leaning or ideology, potentially influencing the reader's political opinion and, ultimately, their voting behavior [10]. There are several approaches to detecting political bias in media, e.g., counting the appearance of certain political parties or ideology-associated words.

Table 1: Media bias tasks and examples

Task	Example from the MBIB datasets
Linguistic bias	"A Trump-loving white security guard with a racist past shot and killed an unarmed Black man during an unprovoked hotel parking lot attack." [48]
Text-level Context Bias	"The governor [...] observed an influx of Ukrainian citizens who want to stay in Russia until the situation normalises in their country" [11]
Reporting-Level Context Bias	In a presidential campaign, one candidate receives disproportionate news coverage. [7]
Cognitive Bias	"Republicans are certain that the more people learn the less they'll like about the Democrats approach" [27]
Hate Speech	"I will call my friends and we go [vulgarity] up that [vulgarity]" [28]
Racial Bias	"black people have a high crime rate therefore black people are criminals" [3]
Fake News	"Phoenix Arizona is the No 2 kidnapping capital of the world" [57]
Gender Bias	"For a woman that is good." [15]
Political Bias	"Generally happy with her fiscally prudent, dont-buy-what-you-cant-afford approach [...]" (classified right) vs "[...] some German voters have also begun to question austerity." (classified left) [27]

3.2 Dataset Collection

Based on our tasks, we select suitable data for each task. These should be available and widely used datasets in line with the guideline by Wang et al. [55]. Furthermore, datasets should be diverse

in the type of bias they cover and have consistent, high-quality annotations. Since no extensive overview of the datasets used in the domain exists to date, we once again assess the list of media bias-related publications mentioned in the literature review by Spinde et al. [43] previously used for the task selection.

In our work, we manually analyze all articles in the list to assess each used dataset, providing a complete overview of utilized datasets. We review 322 media bias-related articles, resulting in a dataset overview of 115 datasets.

Figure 1: The dataset collection and selection process

3.3 Dataset Selection

In benchmarks like SuperGLUE [55], and BigBench [50], one dataset is used per task. Having only a single dataset requires a dataset for every task that is sufficient in size and quality, covering the entire task. For media bias, as described in Section 2.2, no such single dataset per task exists. Either datasets are too small, or they only cover a particular aspect of a task (e.g., datasets associated to the linguistic bias task either cover framing bias or connotation bias but not both). Therefore, we base every task on multiple datasets containing varying definitions of bias, reflecting the dataset-scarce research area as detailed as possible. The variety of initial datasets, which we detail more in Section 4, gives more reason why a benchmark and overview of the domain is required. Mainly, we evaluate our datasets based on the following criteria:

(1) First, we identify whether a dataset is accessible[8] and labeled.

(2) Second, we identify whether the dataset uses the English language. For now, we focus only on English since it is the dominant language in the research domain [42].

(3) Third, we evaluate the dataset size. While bigger datasets usually contain a more balanced range of content, they are often labeled automatically. Smaller datasets are mostly manually labeled and of higher quality. However, they do not exhibit sufficient data points for many current system architectures [48]. We set a minimum of no less than 700 data points per dataset[9].

(4) Fourth, we manually evaluate the dataset quality, in terms of dataset transparency, diversity of sources, overlap with other datasets, and potential annotator training. We summarize the benefits of all chosen datasets within the MBIB repository. We are aware that ultimately, the assessment of quality in MBIB is based on a manual choice, and address this again in Section 6.

(5) Fifth, we require that datasets can be transformed into one unified format as specified in the remainder of this section.

(6) Sixth, datasets need to belong to one of the MBIB's tasks.

Figure 1 includes the number of datasets filtered out based on these criteria. Of the 115 datasets collected, only 74 (65%) are directly available. We discard all 41 others, as well as three which have no labels. Two datasets are not in English. We discard five datasets because they could not be transformed into a unified format.[10] Finally, 15 datasets belong to non-MBIB tasks such as sentiment analysis. After applying the criteria, 22 datasets remain. Table 2 displays the selected datasets for each task with their respective size. Out of the 22 selected datasets two can only be obtained directly from the authors due to copyright issues [14, 27]. MBIB is, therefore, currently available in two versions: in a base and a full module. The base version includes only the datasets directly available and aims to facilitate access. In the full version all 22 datasets are used. We provide guidance on accessing and preprocessing the remaining datasets to create the full version in the MBIB repository. All our experiments (subsection 2.2) use the full version of MBIB.

After the dataset selection, no datasets associated with reporting-level context bias are left. Of the initial six datasets for reporting-level context bias, five are not publicly available, and the only available dataset is not labeled [16]. Reporting-level context bias, therefore, is currently not included in MBIB. We aim to add reporting-level context bias to MBIB as soon as enough data exists.

3.4 Preprocessing

We preprocess all of the above datasets into one unified format consisting of the unique ID of the text segment to be analyzed, an ID indicating to which dataset the statement belongs, the text, a binary label, and, if given, additional labels[11].

While keeping the original labels, we transform all dataset labels into a binary label format. This has three major advantages:

---

[8]We consider a dataset accessible if it is either publicly available, can be directly recreated (e.g., tweet IDs and labels provided), or there is a defined way to obtain it from the authors.

[9]After manual inspection, we conclude that the RacialBias [13] dataset is the smallest set that still included high-quality annotations with sufficient variety in content.

[10]For instance, the dataset provided by [6] contains quotes with associated statements outlining the context, which could not be transformed into a binary label.

[11]We show the exact preprocessing steps for every dataset within the MBIB repository. There, we also show the datasets in more detail.

Table 2: Tasks and datasets in MBIB

Tasks and Datasets	Data Points
Linguistic Bias	**433,677***
Wikipedia NPOV [18]	11,945
BABE [48]	3,673
Wiki Neutrality Corpus [33]	362,991
UsVsThem [19]	6,863
RedditBias [3]	10,583
Media Frames Corpus [22]	37,622
BASIL [9]	1,726
Biased Sentences [25]	842
Cognitive Bias	**2,344,387***
BIGNEWS [27]	2,331,552
Liar Dataset [57]	12,835
Text-Level Context Bias	**28,329***
Contextual Abuse Dataset [53]	26,235
Multidimensional Dataset [11]	2,094
Hate Speech	**2,050,674***
Kaggle Jigsaw [1]	1,999,516
HateXplain [28]	20,148
RedditBias [3]	10,583
Online Harassment Corpus [14]	20,427
Gender Bias	**33,121***
RedditBias [3]	3,000
RtGender [54]	15,351
WorkPlace sexism [15]	1,136
CMSB [37]	13,634
Racial Bias	**2,371***
RedditBias [3]	2,620
RacialBias [13]	751
Fake News	**24,394***
Liar Dataset [57]	12,835
PHEME [60]	5,222
FakeNewsNet [38]	6,337
Political Bias	**2,348,198***
UsVsThem [19]	6,863
BIGNEWS [27]	2,331,552
SemEval [20]	9,783

*Refers to the total number of data points of the task.

(1) It allows an easy combination of different datasets without requiring different model heads.
(2) It follows the task principles set up by Wang et al. [55] to formulate the task as simple as possible.
(3) By keeping the original labels, changes applied to non-binary labels can be tracked transparently[12].

Most datasets already have binary labels. However, we determine a threshold for some datasets with continuous labels to binarize the data. If possible, the authors' recommendation for a threshold is followed. The original format of every dataset is shown in Table 3. For instance, the Kaggle Jigsaw data [1] is labeled on a scale from 0 to 1. The authors recommend using a 0.5 threshold to binarize the label, which we follow for MBIB. Also, we collapse multi-categorical labels into two categories. For instance, for the political bias task, 'right' and 'left' are combined into 'biased' [20, 27].

---

[12]Investigating the non-binary labels in more detail will also be an interesting aspect in future work.

The Liar dataset [57] provides even more labels: 'true', 'mostly-true', 'half-true', 'barely-true', 'false', and 'pants-fire'. The first four labels are combined into a single 'true' label and the last two into one 'false' label. Even though social media-specific elements such as hashtags or emoticons are used in related areas such as sentiment analysis [58], we remove them to not deviate further from a news format. As an additional step, we enrich the FakeNewsNet dataset by scraping the tweets or articles referred to by the Tweet IDs given in the original resource [38][13]. Every decision with regards to the unified format is furthermore detailed in the MBIB repository.

4 DATASET PROPERTIES

Out of the 115 datasets in the dataset overview, 38 contain annotated articles, 32 annotated sentences, 25 annotated social media posts (mainly Tweets), five annotated comments, and two annotated headlines. The most prominent article source for datasets is allsides.com (eight datasets). Five datasets stem from Wikipedia. Ten datasets are created using Amazon Mechanical Turk workers; six refer to other crowd-working platforms. 32 datasets are found to be either self-annotated by the authors or by individually hired annotators. 20 datasets are labeled using distant labeling (retrieving the label for a single article from the outlet's bias score). Overall, we find binary, multi-class, and continuous labels. The data also contains a lot of other annotations, such as bias-inducing words or context data. The datasets have a median of 8,656 data points, with significant variance. The dataset distribution among MBIB's tasks can be found in Section 4.

Figure 2: Dataset distribution over MBIB tasks

Out of the 22 datasets in MBIB, nine contain news articles. Eleven include data from social media, two data from Wikipedia, and one dataset only consists of quotes (one dataset has two different sources). More details on the properties of all datasets can be found in Table 3. The social media datasets use data from Reddit [3, 15, 19, 53, 54] and Twitter [1, 13, 14, 28, 37, 60]. Only [28, 54] further include data from other platforms such as Facebook and Gab[14]. Also, the social media datasets usually focus on specific events [19], or phrases [28, 37]. The Wikipedia-based datasets [18, 33] are both based on Wikipedia's POV label, signaling a potentially biased statement. News articles within the collection come from various widely known sources such as the New York Times [27] and alternative

---

[13]The original dataset does only include IDs, not the texts themselves.
[14]https://gab.com is a microblogging and social networking service.

media sources [57]. Partially the datasets are general collections of news articles [9, 20, 22, 27, 38, 57] and partially they contain articles collected around certain topics [11, 25, 48].

For the annotations, most authors use crowdsourcing [11, 18, 19, 22, 28, 37, 49, 53, 54]. Others train or commission selected annotators [20, 48], annotate themselves [13, 15] or use external annotations [27, 33, 38, 57]. Some contributions further report either an annotator training, instructions, or mechanisms to ensure high annotation quality (such as control questions) [3, 9, 11, 14, 18, 28, 48]. The labels provided range from binary labels [3, 9, 13–15, 17, 28, 33, 38, 48, 53, 60] to multi-class labels [11, 20, 22, 25, 27, 54, 57] and continuous labels [1, 19, 37].

The BigNews Corpus [27], the largest dataset in our collection, is the only dataset not directly labeled on the individual text level. Instead, annotations are based on the outlet's bias label.

5 EVALUATION AND BASELINES

5.1 Evaluation Framework

To evaluate models on MBIB, we introduce a framework defining which metrics should be used and reported. We illustrate the usage of our framework by evaluating five transformer models on MBIB. We use stratified 5-fold-cross-validation on the preprocessed MBIB data, which ensures stable scores while remaining computationally feasible for larger tasks. Also, we balance the classes in each task to ensure an equal representation of classes in each fold and thus ensure unbiased scores.

As the primary performance metric, we choose the F_1-score based on its established usage as a metric in various benchmarks, including those previously discussed [50, 55, 56]. All scores of the five folds are averaged. As the datasets we combine into one task differ in the number of observations they contain, larger datasets can strongly influence the final score. Therefore, to calculate the F_1-scores, we propose two methods:

- The micro average F_1-score: One F_1-score is calculated on the predictions of a model on the entire test set. The scores of the five folds are averaged.
- The macro average F_1-score: Multiple F_1-scores are calculated on the predictions of a model on the test set. One F_1-score is calculated individually for every dataset (from which the data originally stems).

The macro approach ensures that each dataset is represented equally in the final score, regardless of its size.[15] The micro score's simplicity and focus on larger datasets enable an assessment of the impact of dataset size on the model's overall performance when used in conjunction with the macro score. As performed in Super-GLUE [55], we average both scores for all tasks into two (micro and macro) final media bias scores. By reporting the averaged scores, we can evaluate model generalizability. By also providing single-task scores, we can evaluate performances on individual tasks.

5.2 Testing Baselines

As baselines, we test base versions of five available transformer models, ConvBERT, Bart, RoBERTa-Twitter, ELECTRA, and GPT-2,

focusing only on lexical features. [16] Figure 3 displays the performance of the five evaluated models on every task. The results of the best performing models can furthermore be found in Table 4.[17] The

Figure 3: F_1-scores per task

(a) Micro-average

(b) Macro-average

baseline results give a first intuition about the purpose of MBIB as they show that no single transformer model stands out as the best-performing model across all tasks. We find substantial inter-task performance differences. These differences indicate that some tasks, e.g., racial and gender bias, seem easier to detect than other tasks, such as fake news or cognitive bias. However, the performances are very similar within individual tasks. GPT-2 underperforms on most tasks. The distinction between the micro and macro F1-scores highlights a more comprehensive evaluation of a model's ability to detect media bias, revealing its specific strengths and weaknesses. However, it is crucial to explore alternative metrics and delve deeper into the analysis of individual task scores, as further research in this area remains necessary. When creating MBIB, a key concern was that the varying size of the datasets in a task could disproportionately affect the performance of the models. Therefore, the performance on individual datasets is considered in relation to the size of the datasets (Appendix A.2). A positive linear relationship could be expected if there were a direct correlation between size and performance. However, this cannot be observed.

We foresee advancements in the performance on the MBIB and encourage continuous assessment of innovative and refined techniques to maintain this progress.

---

[15]Which is particularly important due to the immense size differences between datasets, which likely skew the macro average F_1-scores towards our more extensive datasets.

[16]So, for instance, for fake news detection, we do not use fact-checking databases. We provide details on these models and why we choose them in the MBIB repository

[17]We show the average final scores within the MBIB repository.

Table 3: Details of MBIB's datasets

Dataset Name	Feature Level	Feature Source	Label Categories	Label Source
Wikipedia NPOV [18]	statements	Wikipedia	binary bias label	Wikipedia editors
BASIL [9]	event spans	news articles	binary bias label	trained annotators
BABE [48]	sentences	news articles	binary bias label	trained annotators
PHEME [60]	tweets	Twitter	binary rumor and veracity label	journalists
Multidimensional Dataset [11]	sentences	news articles	labeled on three bias dimensions	crowdsourcing, expert control
FakeNewsNet [38]	sentences	news articles	binary veracity label	fact-checking websites
Wiki Neutrality Corpus [33]	sentences	Wikipedia	binary bias label	Wikipedia editors
SemEval [20]	sentences	news articles	multi-categorical hyperpartisan label	three annotators
Media Frames Corpus [22]	sentences	news articles	neutral/pro/anti	crowdsourcing
Biased Sentences Dataset [25]	sentences	news articles	multi-categorical bias label	crowdsourcing
Kaggle Jigsaw [1]	comments	Twitter	continuous toxicity label	annotators
UsVsThem [19]	comments	Reddit	continuous bias label	crowdsourcing
BIGNEWS [27]	sentences	news articles	left/neutral/right	allsides.com
Liar Dataset [57]	statements	politifacts.com	multi-categorical veracity label	expert annotators
RedditBias [3]	comments	Reddit	binary bias label	trained annotators
Contextual Abuse Dataset [53]	posts, comments	Reddit	binary abuse label	expert annotators
Online Harassment Corpus [14]	tweets	Twitter	binary harassment label	trained annotators
HateXplain [28]	sentences	Twitter and Gab	binary hatespeech label	crowdsourcing
RtGender [54]	sentences	Facebook, Reddit, TED, Fitocracy	multi-categorical gender perception label	crowdsourcing
WorkPlace sexism [15]	sentences	news articles, quotes	binary gender bias label	trained annotators
CMSB [37]	tweets	Twitter	continuous sexism scales	crowdsourcing
RacialBias [13]	tweets	Twitter	binary racial bias label	annotators

6 DISCUSSION

Already in 2021, Spinde et al. [48] emphasized that media bias systems with better generalization capabilities are needed. The authors concluded that a way to promote such generalization capabilities would be a more refined evaluation scheme, as presented in Check-List [35]. MBIB, for the first time in the media bias domain, provides such a refined evaluation scheme and promotes the development of models with stronger generalizability. Also, MBIB illustrates that many different tasks exist in the media bias domain, and creates awareness of media bias being a complex construct in total.

While the media bias framework of [43] helps to capture this complexity, it has the downside of tasks not being intuitive to understand, sometimes causing unclarity about where a specific bias fits in. Future work should, therefore, improve taxonomies and conceptual overviews of the research domain. The amalgamation of various bias tasks presents a considerable challenge due to factors such as legal disparities, regional variations, and cultural contexts. Addressing these complexities necessitates meticulous dataset curation, annotation guidelines, and a thoughtful approach to evaluation procedures. Nevertheless, it is through this comprehensive combination that we can foster the development of methodologies capable of addressing multiple types of biases concurrently. For

the selection of datasets, we aim to present a stringent justification. However, other compositions of datasets are also conceivable. While some areas in the domain exhibit multiple datasets, others are scarce or have no particular datasets targeting the respective subconcept. The lack of datasets for the reporting-level context bias task additionally limits MBIB. Also, the text-level context and racial bias tasks can be strengthened by more available datasets as they currently only contain two datasets. Especially for text-level context this leaves the informational bias aspect (as described by [43]) uncovered. We hope that our comprehensive dataset overview can promote the creation of further datasets by facilitating the identification of areas with low data coverage. The large amount of social media data within our benchmark reflects that bias on social media is potentially stronger [13]. However, to address social media and news outlets equally, both important sources of information, we propose to focus on developing more news article-related media bias datasets. Additionally, all datasets exhibit some tradeoff between manual labels being expensive, and automated labels not always being precise. Big datasets such as BIGNEWS [27], therefore, likely introduce a lot of noise due to less precise labels. Generally, annotator training might help to increase agreement among the labels [48], while quality control measures such as monitoring annotator performance might enhance overall data

quality [11]. We acknowledge the importance of privacy and GDPR compliance in handling data from social media and online sources. All datasets included in MBIB are anonymized. To ensure continued compliance, we commit to periodically updating the benchmark to reflect changes in the source datasets, maintaining the integrity and relevance of our benchmark while respecting privacy regulations.

In future work on MBIB, other biases (like framing or religious bias) can be considered; even more, we envision the creation of a benchmark that also includes related concepts, such as sentiment, to continuously extend our benchmark to measure any opinion expressed in texts. A continued discussion about which tasks capture media bias best remains necessary to build such a future collection. Additionally, in future work, we will include multi-categorical or numerical data and integrate multiple languages into our benchmark [41], which is yet only focusing on English. We want to increase the overall diversity of MBIB tasks so that it continues to address the complexity of media bias detection in more and more detail.

7 CONCLUSION

This work proposes MBIB, the first ever multi-task benchmark for media bias. MBIB is organized over nine tasks and consists of 22 datasets, which we curate from a list of 115 datasets in total. We evaluate the datasets based on their focus, size, availability, and label quality to filter the original list. We also present a framework showing how future models can be evaluated using our benchmark. By introducing MBIB, we aim to capture media bias as extensively as possible and to give a complete overview of currently available resources in the domain. We believe that MBIB offers a new common ground for research in the domain, especially given the rising amount of (research) attention directed towards media bias. We will continue to update MBIB in the future, and add potential new datasets and tasks.

A APPENDIX

A.1 Model Baseline Results

Table 4: Best average scores per task

(a) Micro-scores

Bias Type	Model	Micro-Score
Linguistic Bias	ConvBERT	0.7126
Cognitive Bias	ConvBERT	0.7044
Text-Level Context Bias	ConvBERT	0.7697
Hate Speech	RoBERTa-Twitter	0.8897
Gender Bias	RoBERTa-Twitter	0.8334
Racial Bias	ConvBERT	0.8772
Fake News	Bart	0.6811
Political Bias	ConvBERT	0.7041

(b) Macro-scores

Bias Type	Model	Macro-Score
Linguistic Bias	Bart	0.7664
Cognitive Bias	ConvBERT	0.4995
Text-Level Context Bias	ConvBERT	0.7532
Hate Speech	Bart	0.7310
Gender Bias	ELECTRA	0.8211
Racial Bias	ELECTRA	0.6170
Fake News	RoBERTa-Twitter	0.7533
Political Bias	ConvBERT	0.7110

A.2 Model Performance on the Dataset Level

Figure 5: F_1-scores per dataset and the size of the testset

REFERENCES

[1] Jigsaw/Conversation AI. 2019. Jigsaw unintended bias in toxicity classification. https://www.kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification/data

[2] 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. [n. d.]. ExT5:

Towards Extreme Multi-Task Scaling for Transfer Learning. abs/2111.10952 ([n. d.]). arXiv:2111.10952 https://arxiv.org/abs/2111.10952

[3] Soumya Barikeri, Anne Lauscher, Ivan Vulić, and Goran Glavaš. 2021. 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)*. Association for Computational Linguistics, Online, 1941–1955. https://doi.org/10.18653/v1/2021.acl-long.151

[4] Eric Baumer, Elisha Elovic, Ying Qin, Francesca Polletta, and Geri Gay. [n. d.]. 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* (Denver, Colorado, 2015). Association for Computational Linguistics, 1472–1482. https://doi.org/10.3115/v1/N15-1171

[5] Camiel J. Beukeboom and Christian Burgers. 2017. Linguistic Bias. In *Oxford Research Encyclopedia of Communication*. Oxford University Press. https://doi.org/10.1093/acrefore/9780190228613.013.439

[6] Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. 2019. Finding Microaggressions in the Wild: A Case for Locating Elusive Phenomena in Social Media Posts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 1664–1674. https://doi.org/10.18653/v1/D19-1176

[7] Dave D'Alessio and Mike Allen. 2000. Media Bias in Presidential Elections: A Meta-Analysis. *Journal of Communication* 50, 4 (Dec. 2000), 133–156. https://doi.org/10.1111/j.1460-2466.2000.tb02866.x

[8] Jo Ellen Fair. 1993. War, Famine, and Poverty: Race in the Construction of Africa's Media Image. *Journal of Communication Inquiry* 17, 2 (July 1993), 5–22. https://doi.org/10.1177/019685999301700202 Publisher: SAGE Publications Inc.

[9] 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. https://doi.org/10.48550/arXiv.1909.02670 arXiv:1909.02670 [cs].

[10] Stanley Feldman. 2013. Political ideology. In *The Oxford handbook of political psychology, 2nd ed.* Oxford University Press, New York, NY, US, 591–626.

[11] 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*. ACM, Virtual Event Ireland, 3007–3014. https://doi.org/10.1145/3340531.3412876

[12] Alan S Gerber, Dean Karlan, and Daniel Bergan. 2009. Does the Media Matter? A Field Experiment Measuring the Effect of Newspapers on Voting Behavior and Political Opinions. *American Economic Journal: Applied Economics* 1, 2 (March 2009), 35–52. https://doi.org/10.1257/app.1.2.35

[13] Torumoy Ghoshal. 2018. Racial Bias Twitter. https://github.com/tgh499/racial_bias_twitter

[14] 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*. ACM, Troy New York USA, 229–233. https://doi.org/10.1145/3091478.3091509

[15] Dylan Grosz and Patricia Conde-Cespedes. 2020. Automatic Detection of Sexist Statements Commonly Used at the Workplace. In *Trends and Applications in Knowledge Discovery and Data Mining*, Wei Lu and Kenny Q. Zhu (Eds.). Vol. 12237. Springer International Publishing, Cham, 104–115. https://doi.org/10.1007/978-3-030-60470-7_11

[16] Valentin Hofmann, Xiaowen Dong, Janet B. Pierrehumbert, and Hinrich Schütze. 2021. Modeling Ideological Salience and Framing in Polarized Online Groups with Graph Neural Networks and Structured Sparsity. https://doi.org/10.48550/ARXIV.2104.08829

[17] 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*. ACM, Melbourne VIC Australia, 195–203. https://doi.org/10.1145/3289600.3291018

[18] 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*. ACM, Melbourne VIC Australia, 195–203. https://doi.org/10.1145/3289600.3291018

[19] 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*. Association for Computational Linguistics, Online, 1921–1945. https://doi.org/10.18653/v1/2021.eacl-main.165

[20] Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, David Corney, Payam Adineh, Benno Stein, and Martin Potthast. 2018. Data for PAN at SemEval 2019 Task 4: Hyperpartisan News Detection. https://doi.org/10.5281/ZENODO.1489920 Version Number: Training and validation v1 Type: dataset.

[21] 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)* (2022-06-01). Cologne, Germany. https://doi.org/10.1145/3529372.3530932

[22] Haewoon Kwak, Jisun An, and Yong-Yeol Ahn. 2020. A Systematic Media Frame Analysis of 1.5 Million New York Times Articles from 2000 to 2017. In *12th ACM Conference on Web Science*. ACM, Southampton United Kingdom, 305–314. https://doi.org/10.1145/3394231.3397921

[23] Lesley Lavery. 2013. Gender Bias in the Media? An Examination of Local Television News Coverage of Male and Female House Candidates: Gender Bias in the Media. *Politics & Policy* 41, 6 (Dec. 2013), 877–910. https://doi.org/10.1111/polp.12051

[24] Nayeon Lee, Yejin Bang, Andrea Madotto, and Pascale Fung. 2021. Mitigating Media Bias through Neutral Article Generation. *CoRR* abs/2104.00336 (2021). https://doi.org/10.48550/arXiv.2104.00336

[25] Sora Lim, Adam Jatowt, and Y Masatoshi. 2020. Creating a dataset for fine-grained bias detection in news articles *(12)*. 1–35.

[26] 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. (2019). https://doi.org/10.48550/ARXIV.1907.11692 Publisher: arXiv Version Number: 1.

[27] Yujian Liu, Xinliang Frederick Zhang, David Wegsman, Nick Beauchamp, and Lu Wang. 2022. POLITICS: Pretraining with Same-story Article Comparison for Ideology Prediction and Stance Detection. (2022). https://doi.org/10.48550/ARXIV.2205.00619 Publisher: arXiv Version Number: 1.

[28] 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 (May 2021), 14867–14875. https://doi.org/10.1609/aaai.v35i17.17745

[29] Seong-Jae Min and John C. Feaster. 2010. Missing Children in National News Coverage: Racial and Gender Representations of Missing Children Cases. *Communication Research Reports* 27, 3 (Aug. 2010), 207–216. https://doi.org/10.1080/08824091003776289 Publisher: Routledge _eprint: https://doi.org/10.1080/08824091003776289.

[30] Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2020. Hate speech detection and racial bias mitigation in social media based on BERT model. *PloS one* 15, 8 (2020), e0237861. https://doi.org/10.1371/journal.pone.0237861

[31] Raymond S. Nickerson. 1998. Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology* 2, 2 (June 1998), 175–220. https://doi.org/10.1037/1089-2680.2.2.175

[32] Christopher Potts, Zhengxuan Wu, Atticus Geiger, and Douwe Kiela. 2021. DynaSent: A Dynamic Benchmark for Sentiment Analysis. 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)*. Association for Computational Linguistics, Online, 2388–2404. https://doi.org/10.18653/v1/2021.acl-long.186

[33] Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2019. Automatically Neutralizing Subjective Bias in Text. (2019). https://doi.org/10.48550/ARXIV.1911.09709 Publisher: arXiv Version Number: 3.

[34] 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)*. Association for Computational Linguistics, Sofia, Bulgaria, 1650–1659. https://aclanthology.org/P13-1162

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

[36] Yasmim Mendes Rocha, Gabriel Acácio de Moura, Gabriel Alves Desidério, Carlos Henrique de Oliveira, Francisco Dantas Lourenço, and Larissa Deadame de Figueiredo Nicolete. 2021. The impact of fake news on social media and its influence on health during the COVID-19 pandemic: a systematic review. *Journal of Public Health* (Oct. 2021). https://doi.org/10.1007/s10389-021-01658-z

[37] 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. (2020). https://doi.org/10.48550/ARXIV.2004.12764

[38] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media. *Big Data* 8, 3 (June 2020), 171–188. https://doi.org/10.1089/big.2020.0062

[39] Vivek K. Singh, Mary Chayko, Raj Inamdar, and Diana Floegel. 2020. Female librarians and male computer programmers? Gender bias in occupational images

on digital media platforms. *Journal of the Association for Information Science and Technology* 71, 11 (Nov. 2020), 1281–1294. https://doi.org/10.1002/asi.24335

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

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

[42] Timo Spinde, Felix Hamborg, and Bela Gipp. 2020. Media Bias in German News Articles: A Combined Approach. In *ECML PKDD 2020 Workshops*, Irena Koprinska, Michael Kamp, Annalisa Appice, Corrado Loglisci, Luiza Antonie, Albrecht Zimmermann, Riccardo Guidotti, Özlem Özgöbek, Rita P. Ribeiro, Ricard Gavaldà, João Gama, Linara Adilova, Yamuna Krishnamurthy, Pedro M. Ferreira, Donato Malerba, Ibéria Medeiros, Michelangelo Ceci, Giuseppe Manco, Elio Masciari, Zbigniew W. Ras, Peter Christen, Eirini Ntoutsi, Erich Schubert, Arthur Zimek, Anna Monreale, Przemyslaw Biecek, Salvatore Rinzivillo, Benjamin Kille, Andreas Lommatzsch, and Jon Atle Gulla (Eds.). Springer International Publishing, Cham, 581–590.

[43] Timo Spinde, Smi 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. *CSUR* (2023). [in review].

[44] Timo Spinde, Christin Jeggle, Magdalena Haupt, Wolfgang Gaissmaier, and Helge Giese. 2022. How do we raise media bias awareness effectively? Effects of visualizations to communicate bias. *PLOS ONE* 17, 4, 1–14. https://doi.org/10.1371/journal.pone.0266204

[45] Timo Spinde, Christina Kreuter, Wolfgang Gaissmaier, Felix Hamborg, Bela Gipp, and Helge Giese. 2021. Do You Think It's Biased? How To Ask For The Perception Of Media Bias. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL)* (2021-09-01). https://doi.org/10.1109/JCDL52503.2021.00018

[46] Timo Spinde, David Krieger, Manu Plank, and Bela Gipp. 2021. Towards A Reliable Ground-Truth For Biased Language Detection. In *Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)* (2021-09-01). Virtual Event. https://doi.org/10.1109/JCDL52503.2021.00053

[47] Timo Spinde, Jan-David Krieger, Terry Ruas, Jelena Mitrović, Franz Götz-Hahn, Akiko Aizawa, and Bela Gipp. 2022. Exploiting Transformer-based Multitask Learning for the Detection of Media Bias in News Articles. In *Proceedings of the iConference 2022* (2022-03-04). Virtual event. https://doi.org/10.1007/978-3-030-96957-8_20

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

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

[50] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, and et al. Shoeb. 2022. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. (2022). https://doi.org/10.48550/ARXIV.2206.04615 Publisher: arXiv Version Number: 2.

[51] Edson C. Tandoc Jr. 2019. The facts of fake news: A research review. *Sociology Compass* 13, 9 (2019), e12724. https://doi.org/10.1111/soc4.12724 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/soc4.12724.

[52] Daphne Joanna Van der Pas and Loes Aaldering. 2020. Gender Differences in Political Media Coverage: A Meta-Analysis. *Journal of Communication* 70, 1 (02 2020), 114–143. https://doi.org/10.1093/joc/jqz046 arXiv:https://academic.oup.com/joc/article-pdf/70/1/114/34053729/jqz046.pdf

[53] Bertie Vidgen, Dong Nguyen, Helen Margetts, Patricia Rossini, and Rebekah Tromble. 2021. Introducing CAD: the Contextual Abuse Dataset. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 2289–2303. https://doi.org/10.18653/v1/2021.naacl-main.182

[54] Rob Voigt, David Jurgens, Vinodkumar Prabhakaran, Dan Jurafsky, and Yulia Tsvetkov. 2018. RtGender: A Corpus for Studying Differential Responses to Gender. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), Miyazaki, Japan. https://aclanthology.org/L18-1445

[55] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. In *Advances in Neural Information Processing Systems*, Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/hash/4496bf24afe7fab6f046bf4923da8de6-Abstract.html

[56] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. http://arxiv.org/abs/1804.07461 arXiv:1804.07461 [cs].

[57] William Yang Wang. 2017. "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)*. Association for Computational Linguistics, Vancouver, Canada, 422–426. https://doi.org/10.18653/v1/P17-2067

[58] Shreyas Wankhede, Ranjit Patil, Sagar Sonawane, and Prof. Ashwini Save. 2018. Data Preprocessing for Efficient Sentimental Analysis. In *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*. 723–726. https://doi.org/10.1109/ICICCT.2018.8473277

[59] Savvas Zannettou, Mai Elsherief, Elizabeth Belding, Shirin Nilizadeh, and Gianluca Stringhini. 2020. Measuring and Characterizing Hate Speech on News Websites. In *12th ACM Conference on Web Science (WebSci '20)*. Association for Computing Machinery, New York, NY, USA, 125–134. https://doi.org/10.1145/3394231.3397902

[60] Arkaitz Zubiaga, Maria Liakata, and Rob Procter. 2017. Exploiting Context for Rumour Detection in Social Media. In *Social Informatics*, Giovanni Luca Ciampaglia, Afra Mashhadi, and Taha Yasseri (Eds.). Vol. 10539. Springer International Publishing, Cham, 109–123. https://doi.org/10.1007/978-3-319-67217-5_8 Series Title: Lecture Notes in Computer Science.