Improving Media Bias Detection with state-of-the-art Transformers

Master Thesis

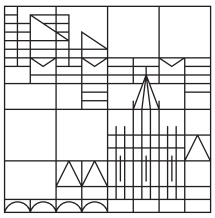
presented

by

Martin Wessel

at the

Universität Konstanz



Faculty of Politics, Law and Economics
Department of Politics and Public Administration

1. Evaluated by	Prof. Dr. Juhi Kulshrestha
2. Evaluated by	Prof. Dr. Bela Gipp

Konstanz, 2023

Acknowledgments

Many people have supported me in the writing of this thesis. Most importantly, I would like to thank Prof. Isao Echizen and Prof. Akiko Aizawa for making it possible to conduct research for this thesis at the National Institute of Informatics in Tokyo. They and their teams have been incredibly welcoming and supportive. Furthermore, I would like to thank Timo Spinde and Terry Ruas from Prof. Bela Gipp's team for their constant guidance and input. My special gratitude to Timo for initializing and organizing our research endeavor. My thanks also go out to the rest of the Media Bias Team, namely Jérôme, Felix, Tomáš, Anna, Fabian, Ann-Christin, Jérôme, and Smi, with whom it was a joy to work. I want to thank Prof. Juhi Kulshrestha for her valuable guidance in the planning of the thesis. Finally, I would like to express my gratitude to the Friedrich-Ebert-Stiftung, namely Renia Doerr, for supporting my research stay in Japan.

Martin Wessel

Improving Media Bias Detection with state-of-the-art Transformers

Abstract

This thesis introduces MBIB, the Media Bias Identification Benchmark. MBIB, inspired by GLUE [Wang et al., 2019b], consists of nine unified media bias tasks and associated datasets. It allows for comprehensive performance analyses and the comparison of models aimed at detecting media bias. An extensive overview of existing media bias datasets is created. Out of this overview of 115 datasets 22 datasets are selected, preprocessed, and combined to form the data basis for MBIB. A framework is developed to evaluate models on MBIB in a unified way. The framework is then used to evaluate transformer models on the benchmark and to set baseline performances. With MBIB this thesis presents a comprehensive and demanding task collection, aimed at developing advanced methods for detecting media bias. Additionally, it shows that the transformer model choice matters less for performance than initially presumed. Finally, it can function as a catalog of current datasets and provide a deeper understanding of remaining research gaps related to media bias.

Table of Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

We are constantly bombarded with news and media coverage. There has been a shift from consuming news from a limited number of TV channels and subscribed newspapers towards online resources [Kitchens et al., 2020]. Articles are being read and widely available on social media, search engines, and other platforms. This shift makes it increasingly difficult for readers to identify the trustworthiness and objectivity of the media they are consuming. A collective term that describes when media presents information that favors a certain political viewpoint or ideology is media bias [Hamborg et al., 2019]. There are diverse reasons for the emergence of media bias. One is the author or producer of an article trying to sway the reader towards their own political or ideological opinion [Hamborg et al., 2019]. News outlets trying to adjust their reporting towards the expectation of their readers for economic benefits [Saez-Trumper et al., 2013] might constitute another cause.

Media bias can have a wide-ranging impact by spreading false or misleading information, which can result in the public forming opinions based on incomplete or incorrect facts [Zaller, 1992]. Not only does this hinder the forming of independent opinions but when recognized, it also endangers the trust of readers towards news in general [Ardèvol-Abreu and Gil de Zúñiga, 2017]. Additionally, misrepresentation can lead to certain topics, opinions, or discriminated groups not appearing in the coverage [Min and Feaster, 2010, Singh et al., 2020, Lavery, 2013]. Media bias can furthermore lay the foundation for the emergence of echo chambers, in which consumers are only confronted with news coherent with and confirming their existing belief system [Cinelli et al., 2021].

Because of these various influences, detecting media bias has long been of research interest [Hamborg et al., 2019]. With the rise of electronic media consumption and advances in machine learning techniques, automatic detection has moved into the realm of possibilities. Having a model that automatically detects media bias gives researchers a tool to measure the occurrences of bias. It could also help users to easier identify the trustworthiness of an article, understand how a bias arose, and generally make an informed decision on their news consumption. Finally, such a model could help the producers self-control if a bias is introduced into their work or whether they can reuse information from other news outlets. The latter could be a decisive step to prevent the spreading and replication of misinformation.

Media bias detection remains difficult because media bias can be inflicted at many different stages throughout the production and consumption of a news article Hamborg et al. [2019]. In addition, media bias often involves subtle and complex linguistic and contextual clues that are difficult for humans and machines to recognize [Beukeboom and Burgers, 2017]. Furthermore, media bias can vary depending on the individual, their political beliefs, and the news outlet, making it difficult to develop a generalizable and accurate detection method. Machine learning models require a large amount of labeled data to learn and make accurate predictions [Soekhoe et al., 2016]. With a high-quality dataset, it is easier for models to learn these subtle and complex linguistic and contextual clues indicative of media bias. High-quality datasets also allow researchers to evaluate and compare different media bias detection methods and to set performance baselines.

In media bias research, new models to detect bias have been introduced continuously. Like in many other natural language-related fields, recently, the usage of transformer models has become most prominent [Spinde et al., 2021b]. However, existing research has only utilized singular models on isolated aspects of media bias [Spinde et al., 2021b, Fan et al., 2019, Huang and Lee, 2019]. There is no coherent overview of which model works best and how to compare models for media bias detection. This thesis aims to address this by introducing the Media Bias Identification Benchmark (MBIB). A media bias task collection with associated datasets and an evaluation framework. Setting model performance baselines on MBIB aims to enable researchers to make an informed model choice.
The construction of MBIB is inspired by a series of widely used benchmark datasets and task collections like GLUE [Wang et al., 2019b], SuperGLUE [Wang et al., 2019a] and BIG-Bench [Srivastava et al., 2022]. They are task collections based on existing datasets that standardize performance measurements. To construct a similar benchmark for media bias, nine tasks are chosen to cover the biases falling under the collective term media bias as comprehensively as possible. To find datasets for each task, to my knowledge, the first extensive overview of existing datasets relevant to media bias research is created. From this list of 115 datasets, 22 are chosen to be included in MBIB. Datasets are preprocessed and brought into a uniform shape. To set a first model baseline on MBIB, five transformer models are chosen by an upstream proxy task. These models are then trained and tested on every task.
An evaluation of the model performances finds that though there are performance differences, there is no best model suitable for all media bias tasks. It furthermore allows for an evaluation of how well the models perform on individual datasets and the composition of each task's datasets.

Publishing the overview of available datasets gives researchers an inventory of potential resources, but it also allows for assessing the current availability and quality of datasets in the field. A wide range of datasets is found to concentrate on a few

types of bias. For other types of bias, like reporting-level context bias, there are close to no datasets. This finding serves as an appeal to focus research capabilities toward these fields.
MBIB o□ers extensive training and testing data in a uniform shape, though publication restrictions still hinder making all associated datasets public.

The remainder of the thesis is structured as follows: chapter 2 discusses existing work related to media bias detection, benchmark dataset collections from other fields, and motivates the need for such a benchmark in the field of media bias. chapter 3 and section 3.2 describe the creation of MBIB. chapter 4 until chapter 6 present the creation of model baselines by selecting the most suitable models and testing them on MBIB. Finally, chapter 7 discusses current limitations and an outlook on future steps.

CHAPTER 2
Related Work

2.1 Media Bias

Media bias, as defined by Hamborg et al. [2019] and Spinde et al. [2021c], refers to the "slanted news coverage" that results from journalists intentionally introducing bias into articles. Media bias can be inflicted intentionally or unintentionally [Baumer et al., 2015], and can occur at multiple stages throughout the production and consumption of a news article [Hamborg et al., 2019].
Other definitions of media bias include one used by D'Alessio and Allen [2000], who divide media bias into gatekeeping bias, coverage bias, and statement bias, and a definition by Mullainathan and Shleifer [2002], who divide media bias into spin bias and ideology bias. Gentzkow and Shapiro [2006] and Lee et al. [2021] define media bias as "slanted reporting" that influences the opinions or judgments of readers in a one-sided manner. Media bias can also occur in non-textual content, such as audio or images, but this thesis focuses on text-based bias.
The influence of media bias on readers' perceptions of events can be significant, potentially impacting their opinions [Zaller, 1992] and even their voting behavior [Gerber et al., 2009]. News consumption amplifies this e□ect through digital platforms such as news aggregators [Bui, 2010].

2.2 Automatic Media Bias Detection

Spinde et al. [2022a] divide automated approaches for detecting media bias with computational methods into four categories. The first category includes approaches that use traditional methods such as measuring word occurrence or frequency [Niven and Kao, 2020, Zahid et al., 2020]. Niven and Kao [2020] use word frequency in news articles to measure selection bias in the news of authoritarian states. Zahid et al. [2020] count tweets and polarity rates to calculate scores on coverage and statements of events, which they identify as the main driver for media bias.
The second category consists of machine learning methods, such as logistic regressions [Recasens et al., 2013] or various classifiers [Baumer et al., 2015]. Chen et al. [2020] use a Gaussian Mixture Model as an alternative machine learning approach.
The third category includes neural network-based approaches, such as non-transformer deep learning models [Chen et al., 2020] and transformer-based models [Spinde et al., 2021c].
The final category comprises graph-based approaches, such as the sentence-graph representations used by Guo and Zhu [2022] to incorporate context from adjacent sentences and the entire article. Spinde et al. [2022a] give a more detailed analysis of the contributions in all four categories. They also show that, especially in the last years, the research focus has shifted primarily towards transformer-based approaches. Following Spinde et al. [2021c] and Spinde et al. [2022b], the models compared in this thesis will aim to identify sentence-level bias induced by word choice.

2.3 Similar Approaches in Other Areas

Over the last few years, benchmarks for natural language tasks have gained in popularity and importance by offering a standardized way of comparing how different models and training strategies perform on these tasks. The field has four influential benchmarks: SentEval, GLUE, SuperGLUE, and BIG-Bench. As this work aims to set up a benchmark for the realm of media bias, a closer look at these four can give guiding insights for MBIB's design.

To evaluate and benchmark different sentence representations Conneau and Kiela [2018] publish SentEval. Sentence representations are embedding techniques such as the embeddings produced by GLoVe [Pennington et al., 2014] or Word2Vec [Mikolov et al., 2013] designed to translate words or sentences into a vector representation. SentEval includes datasets for various binary and multi-class classification tasks like sentiment analysis, image retrieval tasks, and natural language inference, aiming to determine whether a given premise sentence entails or contradicts a given hypothesis sentence. Another task in SentEval is similarity detection, aimed at determin-

ing whether two sentences convey the same meaning. The datasets for SentEval consist of a collection of existing datasets that had been frequently used before. Conneau and Kiela [2018] state that taking well-used datasets would increase other researchers' confidence in the benchmark.

Wang et al. [2019b] create the General Language Understanding Evaluation (GLUE) benchmark, a collection of tasks designed to evaluate the performance of NLP systems in a variety of different language understanding tasks. The language understanding tasks covered by the benchmark include, among others, similarity tasks, inference tasks, and question answering. Similar to SentEval, the datasets used to construct each task are well-used from the literature. One of the key features of GLUE is its ability to provide a single, standardized score that can be used to compare the performance of different NLP systems across the different tasks included in the benchmark. This score allows researchers to more easily compare different systems' relative strengths and weaknesses and helps identify areas where further research and development are needed. Additionally, Wang et al. [2019b] evaluate models for a baseline score on each task. They train models on each task individually and multitask models. To ensure a fair model evaluation, Wang et al. [2019b] only published unlabeled test data. Researchers can upload their model's predictions to a website to calculate a score for the predictions. The Natural Language Processing group at Stanford University developed and maintains the website.

As a reaction towards the strong performance increases on natural language tasks with the rise of transformer models like BERT [Devlin et al., 2019] and GPT [Radford et al., 2018] Wang et al. [2019a] publish an improved version of GLUE called SuperGLUE. The results on GLUE got so good (and outperformed human performances) that it is no longer an adequate benchmark. SuperGLUE mainly updates two aspects of GLUE: A wider variety of tasks are introduced, and the difficulty level of tasks is increased. For every task, a human baseline is measured as a performance baseline. The tasks are chosen based on a public collection process where external researchers could submit potential tasks. Interesting are the criteria which every task had to comply with: As SuperGLUE is aimed at natural language understanding, all tasks should test an aspect of this field. Regarding the difficulty, tasks "should be beyond the scope of current state-of-the-art systems, but solvable by most college-educated English speakers" [Wang et al., 2019a, p.4]. To measure the performance, every task should have a well-defined metric. In the case of SuperGLUE, these usually consist of an F_1-Score, ROUGE, or BLEU. All tasks should have public training data available. Finally, tasks should be in a format that is as simple as possible. These principles, except the overall aim of tasks, will be applied later on to the media bias benchmark.

Srivastava et al. [2022] introduce BIG-Bench, a benchmarking tool for natural language processing systems. BIG-Bench is unique from previous benchmarks in

that it includes a wide range of tasks, including document classification, machine translation, and question answering, totaling 204 different tasks. The primary motivation for creating BIG-Bench is to develop an extensive and diverse set of tasks that are yet too difficult for current models to solve completely. Srivastava et al. [2022] also provide human baselines for every task. One of the critical features of BIG-Bench is its ability to generate large amounts of synthetic data that can be used to test the capabilities of NLP systems. This synthetic data is designed to be realistic and challenging and can help researchers evaluate the performance of their systems in a variety of different scenarios.
Tasks are created in a crowdsourced manner, with different authors handing in individual tasks. To be included, tasks must adhere to a strict set of criteria. The criteria for accepting tasks include, among others: the difficulty, not being solvable by memorizing the internet, novelty, size, and the use of computational resources. Tasks must be written in valid code, be easy to read and interpret, and cleanly capture a specific capability of language models. Furthermore, tasks must be beyond the capabilities of current models, not be solvable by looking up strings in model training data, fill a gap in coverage by BIG-Bench, be well-justified, include at least 32 input-output pairs of examples, and use sufficient computational resources [Srivastava et al., 2022]. BIG-Bench has tasks to identify a model's bias (for instance, whether a model exhibits signs of Gender, Racial or Religious Bias). Furthermore, there is a task to identify social bias in models (testing how the model behaves when confronted with different social groups). There are, however, no tasks included that would test the models' capability to detect media bias or one of its subtypes.

Aside from the development of benchmarks in natural language understanding, recent work in sentiment classification has also involved the comparison and benchmarking of models. Pipalia et al. [2020] conduct a comparative analysis of transformer models for sentiment analysis, comparing five state-of-the-art models. They find that XLNet outperformed other models, such as RoBERTa and T5, on sentiment analysis tasks. Farha and Magdy [2021] benchmark transformer models for Arabic sentiment analysis, focusing on the models' training time and computational cost. Guo et al. [2020] compared transformer models for social media text classification, evaluating the performance of three selected models on 25 different datasets. Mathew and Bindu [2020] compared seven other pretrained transformer models on their ability to detect polarity in movie reviews, finding that XLNet performed the best but also had the highest computational complexity.

2.4 Research Gap

The previous section shows that NLP benchmark dataset collections have gained popularity and importance. They offer training data, standardized performance measurements, and insights into the best model for a certain task. Such a bench-

mark does not yet exist for media bias. An extensive overview of the existing datasets in the field is not available. This thesis, therefore, aims to create an overview of existing datasets and use them to create media bias tasks. The tasks collectively form the media bias benchmark, MBIB.

Since the rise of transformers, the availability of different models has increased substantially. Newer models differ in their architecture, training objective, or designated task. Current research has, however, only utilized singular models such as the BERT transformer model [Devlin et al., 2019] for the media bias classification task [Spinde et al., 2021b, Fan et al., 2019, Huang and Lee, 2019]. There is no exhaustive overview and comparison of which transformer model performs best for media bias detection. This thesis, therefore, will follow the creators of GLUE [Wang et al., 2019b] by setting a model performance benchmark on the newly created media bias benchmark tasks and datasets.

Creating curated benchmark datasets will give researchers access to uniformly structured training and testing data on all relevant media bias tasks. Furthermore, it will allow for a comparable performance metric. Setting model benchmarks on each task will give researchers insights into which models and training objectives are preferable and what performances can be expected. Finally, the dataset overview will also shed light on where datasets are needed or the existing data quality needs to be improved.

CHAPTER 3

Creation of MBIB

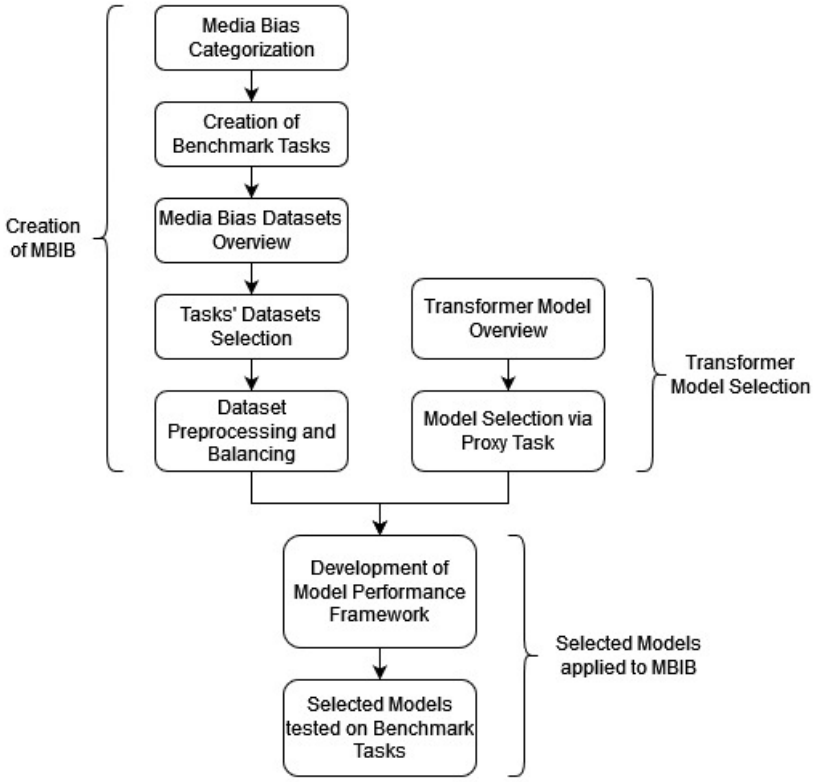


Figure 1. Methodology of the thesis

To create MBIB, first, all relevant media bias tasks need to be collected. Since media bias is an umbrella term under which many di□erent forms of bias appearing in news coverage are collected, this is more complex. Each task should be an independently defined bias under the media bias umbrella term. This requirement facilitates the definition, delimitation, and interpretation of each task. To be selected, the bias needs to fulfill two criteria:

- When occurring in media coverage, the bias constitutes a form of media bias.

- A task is either part of the media bias framework introduced by Spinde et al. [2022a] or is an independent, distinguishable research field of societal importance.[1]

MBIB introduces nine media bias tasks. Four tasks are based on the media bias framework [Spinde et al., 2022a]. These tasks center around how a bias is induced and are to date the only conceptualization aimed at fully covering media bias. Five tasks are independent research fields.
After the task creation, datasets for the tasks needed to be collected. An overview of 115 datasets from in the media bias literature is created. From this overview, a selection of datasets for each task is made based on the datasets' suitability, availability, size, and quality. The data for each task is then preprocessed into a uniform shape and balanced.
After completing the tasks and their datasets, the last thing missing for MBIB is a framework that defines a standardized procedure to evaluate models on each MBIB task. A proxy task is created based on a small but high-quality dataset to filter the best suitable models from the wide availability of existing transformer models. By using the five best-performing models and the developed performance framework baseline performances are set for each MBIB task.
Finally, an in-depth analysis of the results is performed to investigate whether there is one best model choice, how performances di□er on individual datasets, what explains the performances, and what improvements can be made.

3.1 MBIB Tasks

A systematic categorization of media bias is needed to construct a task collection that consists of meaningful tasks which fully cover media bias. That is why four MBIB tasks are based on the framework introduced by Spinde et al. [2022a]. Next to these four tasks, there is a wide range of task candidates constituting independent research areas. Examples of such independent tasks are gender bias or hate speech detection. These tasks constitute media bias when they occur in media coverage but might overlap with the tasks based on the media bias framework. For instance, a linguistic bias might simultaneously induce gender bias. The framework-based tasks already fully cover media bias (as every form of media bias would fall into these tasks). While this framework is new, the independent media bias tasks are well-established in the literature. Including them increases acceptance of the benchmark. Furthermore, these fields are of high societal interest and importance. It is, therefore, relevant to how well models do on these specific tasks.
To find independent tasks candidates the media bias literature collection provided

[1] Note that this criterion implies that the tasks are not necessarily exclusive.

by Spinde et al. [2022a] is assessed for the main types of bias that the research focuses on.[2] The detection of every bias type identified here is considered a potential task for MBIB. After excluding three candidate tasks (reasons for that are discussed at the end of subsection 3.1.2) five tasks remain that complement MBIB.

The tasks introduced by Spinde et al. [2022a] often consist of multiple subtasks. These are taken into account during the dataset collection. To avoid the number of tasks becoming unmanageable and to compensate for the lack of data available for individual subtasks, MBIB only includes higher-level tasks.

3.1.1 Media Bias Framework Tasks

The tasks based on the media bias framework [Spinde et al., 2022a] consist of linguistic, text-level context, reporting-level context, and cognitive bias.

Linguistic Bias constitutes the most researched area of the four. Linguistic bias describes all biases that are induced by lexical features. These features describe which words are used and how they are used to form a sentence [Beukeboom and Burgers, 2017]. Spinde et al. [2022a] divide this task into:
Linguistic intergroup bias, which describes the usage of subtle abstraction forms that deviate objective descriptions to subjective ones so that they reinforce stereotypes between groups [Dragojevic et al., 2017]. An example of linguistic intergroup bias could be using the term "welfare recipients" to refer to people receiving government assistance, as opposed to using the more neutral term "low-income individuals". This subtle abstraction reinforces the stereotype that these individuals are primary beneficiaries when they may be working hard while struggling with poverty.
Framing bias (also called priming) refers to the usage of one-sided words and repeated phrases favorable of one opinion, aimed at guiding the reader's opinion towards it [Entman, 2007]. An example of framing bias could be using the phrase "taxpayer money" to refer to government spending instead of using a more neutral term like "public funds". The phrase can be used by people who are opposed to government spending. Using this phrase, the writer can guide the reader's opinion against government spending.
Similar to framing bias but more subtle, epistemological bias focuses on the usage of verbs or adjectives that induce certain assumptions on the veracity of a statement and the speaker's viewpoint on the matter [Recasens et al., 2013]. An example of epistemological bias could be the phrase "it is widely accepted" when referring to a certain belief or opinion. This phrase implies that the belief is true, even though it may not be supported by evidence or fact. By using the phrase, the speaker tries to induce the assumption that the belief is true and tries to get the listener to accept it.

[2]The same literature collection is used later on to create the overview of media bias datasets.

Bias by semantic properties describes the same as framing bias not by individual words but by sentence structure [Greene and Resnik, 2009]. Bias by semantic properties could be stating that "the unemployment rate is increasing" to describe a situation where the rate of people out of work is growing. The sentence structure implies that the situation is negative and suggests that something needs to be done to address it.
Finally, connotation bias describes the introduction of connotations to alter a statement's meaning. While seemingly an objective statement, the existence of connotations for certain words may completely change the interpretation [Rashkin et al., 2016]. An example of connotation bias could be the phrase "government handouts" to refer to government programs designed to help people in need. This phrase carries a negative connotation and implies that the people receiving the assistance are undeserving.

Text-Level Context Bias acknowledges that a statement rarely stands on its own and puts the focus on how context, the text surrounding a statement, changes the interpretation of the statement. One subtask of text-level context bias is statement bias, which refers to the author introducing his own opinion into a text. Statement bias can already occur through criticizing one side more often than the other [D'Alessio and Allen, 2000]. An example of statement bias is when an author writes a biased article about a particular political candidate. The author might frequently comment negatively about the candidate while praising their opponent. They might also use loaded language to portray the candidate negatively or exaggerate certain facts to make the candidate look bad.
Other subtasks include phrasing bias, context-dependent non-neutral words [Hube and Fetahu, 2019], and spin bias, including unnecessary or excluding necessary information [Mullainathan and Shleifer, 2002]. An example of phrasing bias is when an author uses an inflammatory word such as "scandal" to refer to a much less serious event. An example of spin bias is that an author leaves out that a particular political candidate supported certain legislation to make it appear as though they are against it.

Just like text-level context bias **Reporting-Level Context Bias** is focused on the surroundings of statements. However, not on the surrounding text but the reporting circumstances. One of the most apparent biases in this task is selection bias. It describes bias that arises through decisions made by editors and journalists on what events to report on and what sources to use [D'Alessio and Allen, 2000]. There are limitations on the topics and events that newspapers can cover, but how they choose the most relevant might be influenced by multiple factors, e.g., by what they think the audience wants to read or personal preferences. An example of selection bias in the media can be seen in the 2016 US presidential election. Before the election, many major news outlets declared support for one of the candidates, and much of the coverage is focused on events that favored that candidate [Patterson,

2016]. This type of media bias can have a powerful influence on public opinion and shape election outcomes.
Like selection bias, coverage bias focuses on how balanced the news production process takes place. Coverage bias occurs when one side is disproportionately presented more than the other sides [D'Alessio and Allen, 2000]. This bias can be within a single article and the number of articles released for a certain topic or opinion. One example of coverage bias in Europe can be seen in the media coverage of the European migrant crisis. Here British media overwhelmingly reported negatively about the events compared to outlets in other European countries [Bennett et al.].
Proximity bias describes a bias occurring when news outlet favor covering events that happen close by. Either geographically by preferring local or national news or predominantly iterating opinions close to those of the readers [Saez-Trumper et al., 2013]. For example, a local news outlet may focus on a crime that happened in their city while ignoring a similar crime that happened in another city.

Cognitive Bias focuses on the bias induced by the reader's perception of news. Like on the reporting level, readers might introduce bias by reading articles from outlets and sources based on their worldview. Media consumers might also be more prone to believe sources that approve of their opinions and choose to in the future only consume articles from these sources, making the effect self-reinforcing [Nickerson, 1998].
According to Spinde et al. [2022a], cognitive bias can be divided into selective exposure and partisan bias. Selective exposure focuses on consumers choosing which articles to read and aligning them with priorly formed opinions [Spinde et al., 2022a]. Social media can amplify this effect, as friends and like-minded acquaintances share articles. One example of selective exposure is a reader only following news outlets that align with their political beliefs. For instance, if a person identifies as progressive, they might follow news outlets such as Die Zeit, taz, or the Guardian instead of Fox News, Die Welt, or The Daily Mail. By only following outlets that agree with their preexisting beliefs, they are less likely to be exposed to news that would challenge their beliefs and reinforce their existing biases.
Partisan bias describes the effect of readers being more likely to believe an article is true if it is consistent with their worldview. Vice versa, they are also more likely to dismiss the veracity of articles with opposing views [Gawronski, 2021]. An example of partisan bias can be seen in how people interpret the same information. Depending on their political beliefs, people may interpret the same facts differently and come to different conclusions. For example, when presented with the same economic data, one may argue that it is evidence of a strong economy. In contrast, another individual may argue that it is evidence of a weak economy.

3.1.2 Independent MBIB tasks

The following types of bias (hate speech, fake news, racial bias, gender bias, and political bias) constitute a form of media bias when occurring in news articles. However, they have developed independently as areas of research. Reasons for this might be a big societal interest in these specific topics and often clear identification. Because of this relevance, including the following five tasks into MBIB makes the benchmark more meaningful. The tasks are, however, not exclusive from the media bias framework tasks. In fact, they are often induced by the above-described concepts. Since the independent tasks are often easier to apprehend and make out, they can be valuable in explaining why a statement is biased and why the bias might have harmful impacts.

Most researchers agree that **Hate Speech** contains directing toward a specific target a threat of violence or hate [Fortuna and Nunes, 2018]. Definitions differ on whether offensive language already constitutes such a case [Mathew et al., 2021]. After comparing multiple definitions from the literature [Fortuna and Nunes, 2018, p.85:5] conclude that hate speech is a "language that attacks or diminishes, that incites violence or hate against groups, based on specific characteristics such as physical appearance, religion, descent, national or ethnic origin, sexual orientation, gender identity or other, and it can occur with different linguistic styles, even in subtle forms or when humor is used." Examples of hate speech are, e.g., from the HateXplain dataset [Mathew et al., 2021]: "I will call my friends and we go [...] up that [...]", or from Golbeck et al. [2017]'s online harassment corpus: "[...] and [...] are abominations. The [...] love them". Soral et al. [2018] find that hate speech through desensitization increases prejudice and influences intergroup perception.
While commonly occurring on social media in the news domain, it is much rarer. This is also mirrored in the availability of datasets that exclusively come from non-news domains. Due to its damaging impact, however, being able to detect hate speech in the news remains relevant. Zannettou et al. [2020] expand the detection to comments posted under news articles where hate speech is common. Since the comments are often consumed simultaneously with the articles, this increases the relevance of hate speech detection in media. Mozafari et al. [2020] fine-tune a BERT transformer model [Devlin et al., 2019] to build a hate-speech classifier by utilizing some of the datasets also included in MBIB and find that it significantly outperforms previous approaches.
Hate speech detection is included as an independent media bias task because news organizations need to pay special attention not to amplifying hate speech spreading on social media. When occurring, it creates unsafe and hostile environments for the targeted groups while not being limited to any group, region, or ideology.

Fake News describes published content based partly or completely on false claims and premises. This false content is presented as being true to deceive the reader

[Tandoc Jr., 2019]. In its intent to deceive the reader, it differs from parodies or fiction, which might fulfill the same criteria but are recognizable as such. Fake news has increasingly been used in a politicized context to defame opposing news outlets [Tandoc Jr., 2019]. Social media is often cited as the main distribution platform for fake news as there is not one central responsible news outlet, but users can share faked news articles with each other.
Fake news can have serious consequences. For instance, Rocha et al. [2021] found that fake news concerning the Covid-19 pandemic led to severe psychological disorders. With external and context knowledge, fake news can be easier to identify. For example, "Phoenix Arizona is the No 2 kidnapping capital of the world" or "Under the Iran nuclear deal we give them 150 billion we get nothing" from the liar dataset by Wang [2017] can only be identified as false when fact-checking it.
Research on fake news detection is often based on two strategies: Detecting fake news through linguistic features or comparing content to verified information. Verifying news by comparing it to a ground truth database is an obvious yet labor-intensive strategy. When setting model baselines in chapter 6, fake news will solely be detected through linguistic features, as this strategy is applied to all tasks. However, other approaches, e.g., fact-checking events against a database, are possible on MBIB.
Other types of bias, such as spin bias, might also use misleading information. However, fake news is distinct in its deceiving intent. Additionally, it is tough to identify. Combined with fake news being a globally occurring problem with serious consequences, it makes it an important task to be tackled by media bias-detecting methods.

Chiazor et al. [2021] propose that in the context of media, systemic **Racial Bias** can be found in "news that attempts to portray in a negative light any minority group more than it portrays in a negative light any majority group and vice versa". This bias is often identified by the wording used [Adegbola et al., 2018] and induced by linguistic or text-level context bias. Examples of these kinds of racially biased statements are, e.g., "black people have a high crime rate therefore black people are criminals" or "black people are just inherently criminals" (RedditBias dataset [Barikeri et al., 2021]). Fair [1993] find, e.g., that with 'Africa' words like 'other,' 'black,' or 'primitive' are overrepresented. However, research also suggests that racial bias in the news (especially regarding Africa) has recently declined [Nothias, 2018]. Racial bias might also influence the selection of events in news coverage. Min and Feaster [2010] find that in the reporting of missing children cases in the US, children with an African American background are disproportionately underrepresented. In the US, numerous studies find that criminals with African American backgrounds and Caucasian victims are overrepresented in news coverage [Min and Feaster, 2010]. Closely connected to reporting-level bias, this might be either caused by news outlets trying to appeal to a certain audience or mirror the journalist's social background. Racial bias in news coverage can severely impact affected minorities, as it can strengthen stereotypes and discrimination [Dukes and Gaither, 2017]. While

openly racist statements in news coverage might be rare, the continuing existence of discrimination through wording or representation and the severe societal impact it can have make racial bias a relevant task to consider when researching media bias. Racial bias is often associated with hate speech. While there can be cases of hate speech based on racist motives, the two tasks are distinct. Some statements are racially biased but not hate speech, for instance, the statements from the Reddit-Bias dataset given earlier. Xia et al. [2020] find that there often is a racial bias in the annotation of hate speech datasets, with African American English being more likely to be annotated as hateful.
There have been numerous methods proposed to detect racial bias in the media. Jacobs et al. [2018] identify racial bias by counting co-occurrences of specified word pairs in news articles. In such an approach, however, semantic nuances of the statements are not captured. Kroon et al. [2021] improve the detection by using word embeddings trained on 3 million Dutch news articles. The word embeddings of relevant words are then analyzed on whether they contain ethnic association. Mozafari et al. [2020] also fine-tune their BERT model to detect racial bias. As racial bias is a bias that has been shown to occur in news coverage is distinct, and can have a severe societal impact it is included in MBIB as an independent task.

Gender Bias describes treating one sex either more favorably or discriminating against it. This discrimination often manifests in the underrepresentation of a gender. Examples of gender bias, induced through framing and word choice, in this case discriminating against women are, e.g., "For a woman that is good." or "Leave running the company up to men" (from the workplace sexism dataset [Grosz and Conde-Cespedes, 2020]).
Gender bias can be induced by all four cognitive bias tasks, linguistic bias through word choice or framing, text-level context bias through sexist phrasing, reporting-level context through underrepresenting a gender in reporting, and, finally, cognitive bias through reader-projected stereotypes. The presence of gender bias in the media can have severe impacts on, for instance, the perception and choice of professions, as some jobs will be perceived not as suitable for a certain gender and role models are missing in the public perception [Singh et al., 2020]. Another impact gender bias can have is on voting decisions through the underrepresentation or discrimination of certain candidates [Lavery, 2013]. The severe societal impact that gender bias can have makes it a relevant additional task for media bias research and MBIB.
In the NLP research area, gender bias is a well-researched field. One string of research focuses on using NLP methods to detect gender bias. Another focuses on gender bias induced through, e.g., language models [Costa-jussà, 2019]. Bolukbasi et al. [2016] investigate the latter by examining word embeddings for their gender characteristics. Based on the direction of these word embedding vectors regarding the word embedding vectors of gendered words (e.g., by a projection of $\overrightarrow{word} \cdot (\overrightarrow{he} - \overrightarrow{she})$), they show that not only are many words associated to one gender. Also, sexist operations with word embeddings are possible. These algorithmic biases are intro-

duced through gender bias present in the training corpus. Based on the direction of word embeddings, Bolukbasi et al. [2016] propose a method to correct gender bias by neutralizing the gender direction of a word embedding. Gonen and Goldberg [2019], however, show for this and other neutralization methods that the changes made only appear to work on the surface while the underlying gender association remains. Also, the transformer models used in this survey will not be free of this underlying bias arising from biased training corpora.
In the research utilizing NLP models to detect gender bias in news articles, Dacon and Liu [2021] conduct a large-scale analysis of news abstracts produced by news recommender systems. They classify each abstract as belonging to a certain gender (or none) by counting male or female possessive nouns in each abstract. Dacon and Liu [2021] then measure the gender distribution over all abstracts and the gender association of certain professions and adjectives. Dacon and Liu [2021] find that not only are women underrepresented in news articles but also stereotypes manifest in words associated with women. Grosz and Conde-Cespedes [2020] use an LSTM deep learning model to detect sexism in workplace contexts automatically. Other fields of research include gender bias in advertisements [Sweeney, 2013] , in law [Pinto et al., 2020], or in educational materials [Raina, 2012].

Political Bias, also referred to as partisan bias, describes a news article having a political leaning or ideology. Political ideologies "operate at the societal level to organize political debate by allowing political parties to o□er more or less coherent policy platforms" [Feldman, 2013, p.591]. Often, these ideologies are connected to specific political stances (e.g., a strong welfare state vs. a minimal state). Typically, the ideologies in news articles are divided into "left" and "right" leaning. An article is biased if it tries to influence the reader in either direction. An example of political bias is "Generally happy with her fiscally prudent, dont-buy-what-you-cant-a□ord approach, German voters are poised on Sunday to give Mrs. Merkel [...] a third full term in power in Berlin." classified "right". Opposing to that stands "Ms. Merkel has softened her stance saying that Germany is open to stimulus to spur growth and some German voters have also begun to question austerity." classified "left" (both examples from the BigNews Corpus by Liu et al. [2022]). Political bias occurs either through lexical features as a form of linguistic bias or is induced by the reader's political opinion as a form of cognitive bias. With the focus on political leaning, the political bias task is distinct from the other media bias tasks. The presence of political bias in media can have a decisive influence on political decisions as, e.g., the media can influence the reader's political opinion and ultimately their voting behavior [DellaVigna and Kaplan, 2007]. With its clear distinction from the other bias task and high societal relevance and research interest, political bias constitutes the last MBIB task.
Many approaches have been proposed to detect political bias in the media. The most simplistic approaches are based on counts of the appearance of certain political parties or ideology-associated words in news articles [Lazaridou and Krestel,

2016] or simply analyzing manually annotated articles [Budak et al., 2016]. To detect political bias automatically, Chen et al. [2020] utilize a Recurrent-Neural-Network to detect political bias in news articles classified by allsides. Sinno et al. [2022] compare multiple deep learning strategies for political bias detection based on a multi-dimensional dataset they curate themselves in which they split the political bias into the topics: economical, social, and foreign. Their fine-tuned BERT-based detection system works best compared to LSTM-based approaches.

There are other candidate tasks that are omitted from MBIB. Two of these are framing effects and group bias. Framing effects refer to how media organizations present information to the reader to influence their perception and understanding of it. This framing is done by emphasizing certain aspects while downplaying or ignoring others to create a particular interpretation [Spinde et al., 2022a, Entman, 2007]. However, it remains unclear how this differentiates itself from the framing bias discussed as a subtask of linguistic bias. Group bias refers to bias toward a specific group. It serves as an umbrella term for gender bias, racial bias, and religious bias. Group bias can have other overlaps with, for instance, hate speech. To avoid overlaps, it is not considered for MBIB. However, it could serve as an alternative task to gender and racial bias. Religious bias, a bias arising when discriminating against a group based on their religion [Hart et al., 1980], is another potential additional media bias task. It is mainly not included due to a lack of research (only one mention in the collection of media bias literature).

3.1.3 Media Bias-Related Concepts

All previously mentioned tasks are subtasks of media bias (if present in media coverage). Apart from these tasks, there are numerous media bias-related tasks. These tasks are closely related to media bias detection but do not necessarily identify forms of media bias. One of these tasks is sentiment analysis: The detection of the sentiment of an article, often classified as “positive” or “negative”. Also, stance detection, the identification of the stance of an article towards a particular topic, and topic detection, the identification of an article’s topic, are media bias-related fields. Not being subtasks of media bias and well-researched areas, they are not considered further in this work. They do, however, have the potential to improve the detection of media bias through multi-task approaches [Spinde et al., 2022b] or by applying findings from these areas to media bias research.

3.2 Dataset Collection, Selection, and Preprocessing

3.2.1 Creation of Media Bias Dataset Overview

After defining tasks for MBIB in section 3.1 the next step in creating MBIB consisted of finding suitable datasets for every task. These datasets should be in line with the task guidelines mentioned in section 2.3 set by Wang et al. [2019a] publicly available and already well used in research practice. Both criteria ensure the acceptance and usage of the chosen datasets later on. Finding suitable datasets for each task is challenging because no extensive overview of the datasets used in the area of media bias exists to date. That is why an overview is created to capture all datasets belonging to the realm of media bias research.
The dataset overview is constructed based on the Spinde et al. [2022a]'s crawl of scientific media bias articles. For this crawl websites, such as "Google Scholar", are crawled for articles containing media bias research-related words. The crawled articles are then manually reviewed, classified, and added to an extensive literature list on media bias [Spinde et al., 2022a]. For the dataset overview, every article from this list is checked on which datasets it uses. The datasets are then reviewed for availability, size, feature level, either article or sentence level, and feature source, e.g., Wikipedia or Allsides.com. In addition, it is recorded whether the datasets are labeled and by what annotation source, e.g., Mechanical Turks or self-annotation. Additionally, the datasets' task classification, the paper they are introduced in, and a dataset description are added to the overview. The task classification of each dataset is based on the task they are used for in the respective article.
322 media bias-related articles are manually checked for the overview, resulting in an overview of over 100 used datasets. Afterwards, a search is conducted to find datasets not mentioned in the articles crawled by Spinde et al. [2022a], resulting in an overview of 115 datasets. The dataset overview is publically available as it is a valuable resource for other researchers to find suitable datasets and to see where data availability is still scarce.[3]

3.2.2 Properties of the Datasets

Though the overview consists of a variety of datasets, the properties of the datasets vary substantially. Most importantly, around 60% of the datasets found are not publicly available. Datasets are retrieved after contacting the authors for some of these deemed of high potential value for MBIB. The main data sources are news sites, Wikipedia, or social media platforms such as Twitter and Reddit. The sizes varied from small, often manually annotated, to large, often automatically anno-

[3]The complete dataset overview created can be found under `https://docs.google.com/spreadsheets/d/1BXcDcnBluSzv1bwwAEpRH61ObXd3Mxf66qsOVxilTXM/edit?usp=sharing`

tated datasets. Automatically labeled datasets retrieve the label for a single article from the outlet's bias score. However, they are always based on a human label (for example, the outlet). Since these approaches rely on a distant human label, here, they are called distantly labeled.
The feature level ranges from individual sentences over paragraphs to entire articles. Labels vary from binary, over multi-class, to continuous labels. The data also contains many other annotations, such as bias-inducing words or context data. Though the majority of datasets are labeled, some of the datasets also only consist of unlabeled data. They are included since they might be helpful in later research involving unsupervised language tasks. Most importantly, the available documentation of datasets di□ered widely. While it is straightforward how the data is stored for most, some completely lack available documentation, making it hard to interpret how and where the data is stored.

3.2.3 Dataset Selection

The next step in the construction of MBIB consists of selecting fitting datasets for each task from the dataset overview. Only if the chosen datasets su□ciently cover a task is the benchmark meaningful.
In benchmarks like SuperGlue [Wang et al., 2019a] and BigBench [Srivastava et al., 2022] one dataset is used per task. A single dataset per task requires having a dataset for every task that is su□cient in size and quality and that covers the entire task. For media bias, no such single dataset per task exists. Either the datasets are too small, only cover a certain aspect of a task, or the data foundation is too far removed from news articles. That is why every task is based on multiple datasets. Only a subset of datasets is chosen (as opposed to using every dataset) to remain computationally feasible and reduce preprocessing expenses. As the tasks are not exclusive and datasets have been used for various tasks, some datasets will be found in multiple tasks.

The dataset selection for MBIB is based on multiple criteria. Most importantly, datasets need to be available and labeled. Furthermore, size and data quality are assessed. Since bigger datasets are proportionally less work to preprocess and allow for a more balanced model training, they are preferred. However, the big datasets are often only labeled distantly, while many of the smaller ones are manually labeled and of higher quality. Therefore, the goal is to construct a mixture of high-quality and large-size datasets. Only very small datasets are discarded immediately. The third criterion is the ability to put all the datasets in one task in a consistent binary shape so that combining datasets is possible.
Reporting Level needs to be excluded from MBIB for now since too few datasets are available. Once su□cient data for Reporting Level Bias is available, it should be added back into MBIB. An overview of the final MBIB tasks and selected datasets

can be found in Table 1.

Table 1. MBIB tasks and datasets

Linguistic Bias		Cognitive Bias		Text Level Context Bias		Hate Speech	
Dataset	Size	Dataset	Size	Dataset	Size	Dataset	Size
Wikipedia NPOV	11,945	BIGNEWS	2,331,552	Contextual Abuse Dataset	26,235	Kaggle Jigsaw	1,999,516
BABE	3,673	Liar Dataset	12,835	Multidimensional Dataset	2,094	HateXplain	20,148
Wiki Neutrality Corpus	362,991					RedditBias	10,583
UsVsThem	6,863					Online Harassment Corpus	20,427
RedditBias	10,583						
Media Frames Corpus	37,622						
BASIL	1,726						
Starbucks	842						
Sum	433,677		2,344,387		28,329		2,050,674

Gender Bias		Racial Bias		Fake News		Political Bias	
Dataset	Size	Dataset	Size	Dataset	Size	Dataset	Size
RedditBias	3,000	RedditBias	2,620	Liar Dataset	12,835	UsVsThem	6,863
RtGender	15,351	Wasseem	7,700	PHEME	5,222	BIGNEWS	2,331,552
WorkPlace sexism	1,136	RacialBias	751	FakeNewsNet	6,337	SemEval	9,783
CMSB	13,634						
Sum	33,121		11,071		24,394		2,348,198

One drawback of the data selection that will also become visible in the next section subsection 3.2.4 is that many of the datasets are not based on news articles but on social media or Wikipedia. As both these sources di□er from news articles in many aspects, this limits MBIB. Though ideally, all datasets would be solely based on news articles, for many tasks little to no such dataset exists. To get at least an approximation of how well the classifiers build for media bias detection do on these tasks, non-news article-based datasets are included.

3.2.4 Datasets

The following section gives a short overview of all 24 selected datasets for the eight MBIB tasks.

3.2.4.1 Linguistic Bias

BABE is introduced by Spinde et al. [2021b]. BABE contains 3,673 manually annotated news article-based sentences and incorporates the MBIC dataset introduced by Spinde et al. [2021a]. It is specifically designed for media bias research and stands out through its thorough annotator training. Annotators are required to have a substantial background in the media bias domain, underwent specific training, and are provided with annotation guidelines. BABE provides binary bias labels and various information on biased words, label opinions, or outlet labels. The BABE dataset has been used previously to set a benchmark for automated media bias detection

[Spinde et al., 2021b].

The **Wikipedia NPOV Corpus** [Hube and Fetahu, 2019] uses Wikipedia's POV tags to collect biased and unbiased sentences, resulting in a collection of 11,945 sentences. If a sentence is flagged with a POV tag, it indicates a violation of Wikipedia's neutral point of view (NPOV) policy. Hube and Fetahu [2019] collect a vast amount of such flagged sentences and filter those where only a single statement is corrected in the revision. Five thousand statements are sampled and annotated by crowdsourcing on whether they entail bias. Only those statements annotated as biased are added to the NPOV Corpus as biased sentences. Finally, the NPOV Corpus is enlarged with neutral statements.

The **Wiki Neutrality Corpus** is similar to the Wikipedia NPOV based on Wikipedia's POV labels and is introduced by Pryzant et al. [2019]. It, however, contains biased statements as well as their neutral corrections. Opposed to the Wikipedia NPOV Corpus, crowdsourcing annotators did not review the statements. No manual reviews allow the corpus to be significantly more extensive. However, since Hube and Fetahu [2019] report that only around a third of flagged statements are found to be biased by the annotators, it also indicates a lower data quality. For the MBIB task, the dataset is split up, and every statement is labeled biased or unbiased based on its POV-flagged statement or correction. The final Neutrality Corpus has 362,991 data entries.

UsVsThem is introduced by Huguet Cabot et al. [2021] and consists of 6,863 Reddit comments annotated for populist attitudes. Comments are collected based on being directed towards one of the following groups: "Immigrants, Refugees, Muslims, Jews, Liberals, and Conservatives" [Huguet Cabot et al., 2021]. Furthermore, comments needed to be a direct response to a news article. Comments are then annotated by MTURK crowdworkers on an Us-vs-them scale identifying a recognizable a□liation to or against a group. Such an a□liation is here considered to be a biased statement. A binary version of the scale is also provided, which is used here. Additionally, Huguet Cabot et al. [2021] also assess the emotions, associated groups and provide di□erent labels for each comment.

The **RedditBias** dataset [Barikeri et al., 2021] is made out of 10,583 annotated Reddit posts. The dataset consists of four subcategories: Religion, race, gender, and queerness. It covers diverse topics, such as news, politics, entertainment, and technology. Out of a collection of more than 1.2 million comments, a representative subset is chosen and annotated by crowdsourcing. Annotators are given training examples and required a high accuracy to be included. In the dataset, sentences and entire phrases are annotated on whether they are biased. Only the sentences (including the phrase) are taken from the dataset for the linguistic bias task. Because of the di□erent subcategories, parts of the dataset can also be used in the

racial and gender bias categories.

In the **Media Frames Corpus** [Kwak et al., 2020] 1.5 million news articles from major English-language news outlets, including the New York Times, CNN, and Fox News, are collected. 37,622 of these articles are annotated regarding "media frames". Media frames refer to how a news story or article is framed or presented to emphasize certain aspects of an issue and downplay or ignore others. Media frames can be used to shape public opinion and influence policy decisions. The text of each article is shortened to 225 words by the authors. The annotation process is done through crowdsourcing. While each article is annotated with multiple frames, these mainly consist of topic descriptions. The label chosen here as a proxy for bias is whether the author is "Neutral" or "Pro/Anti".

BASIL introduced by Fan et al. [2019] contains 1,726 manually annotated biased spans from 300 articles. The dataset includes articles from various sources, including mainstream and alternative media. The articles are collected over several years and span a wide range of topics, including politics, health, science, and technology. Spans are annotated by human annotators trained to detect bias and misinformation in the statements. Annotations are then cross-checked for high inter-annotator agreement. While the annotations contain more detailed information on the type of bias and its direction, for this task, only whether a span is annotated as biased is considered.

The **Starbucks** dataset [Lim et al., 2020] is made of 842 sentences from 46 news articles, manually annotated for bias using crowdsourcing. The news articles are all about four different events. These events cover various topics such as politics, sports, and economics. The 4-5 given annotations from different annotators per sentence are averaged to retrieve a single score. Multiple label classes are concatenated to form a binary label. For instance, the biased and very biased categories are concatenated to only biased. Lim et al. [2020] find that the inter-annotator agreement is relatively low and put forward the hypothesis that this might be based on predefined opinions of annotators already familiar with the events described in the news articles.

3.2.4.2 Cognitive Bias

The **Liar Dataset** [Wang, 2017] contains 12,835 statements with six different labels for the degree of truthfulness of the statements and is specifically designed to evaluate models for detecting fake news. The statements are scraped from politifacts.com but stem from various sources, from fake news websites to reputable sources such as The New York Times, and are manually annotated. The dataset is balanced so that the number of true and false news is similar. For the construction of MBIB, the labels are collapsed to a binary true or false format.

The **BigNews** Corpus created by Liu et al. [2022] contains a crawl of articles classified by the media outlet's political leaning as defined by allsites.com. The 3.689.229 articles are classified as neutral or left/right-leaning. The label for left/right is concatenated to one bias label to binarize it. The articles are split on the sentence level to make the length compatible with the rest of the data. The resulting data, however, is only distantly labeled.

3.2.4.3 Text-Level Context Bias

For the **Contextual Abuse** dataset [Vidgen et al., 2021], the authors collected comments and posts from 116 subreddits. They manually annotated them for various types of abusive or hateful language and the target of the abusive language. The collected comments are annotated by crowdsourcing. Annotators are required to reach a consensus regarding the annotation choice. Annotation conflicts are therefore discussed among annotators until a consensus is reached. Afterward, annotations are reviewed by an expert. After missing text entries are dropped, 26,235 are labeled neutral or abusive.

The **Mulitdimensional Dataset** introduced by Färber et al. [2020] uses crowdsourcing to annotate 2,094 sentences on three bias dimensions. The three bias dimensions are hidden assumptions and premises, subjectivity, and framing. For the cognitive bias task, when a majority of annotators agreed upon either dimension, the sentence is labeled as biased. The sentences stem from 90 di□erent news articles about the Ukraine crisis being categorized as either being pro-West, pro-Russian, or neutral. Annotations are made by crowd-workers who had to answer test questions aimed at ensuring the quality of the annotations.

3.2.4.4 Hate Speech

The **Kaggle Jigsaw** dataset [AI, 2019] consists of 1,999,516 tweets annotated for toxicity. It is published as part of a Google competition to classify toxicity in 2019. Toxicity is given as a continuous variable between 0 and 1. Following the authors, a tweet is considered hate speech if it reaches a threshold of 0.5. According to Google [2022], "toxicity is defined as anything rude, disrespectful, or otherwise likely to make someone leave a discussion".

HateXplain introduced by Mathew et al. [2021] entails 20,148 tweets from Twitter and Gab annotated by Amazon Mechanical Turk (a crowdsourcing platform) for hate speech as well as o□ensive language. Annotations include the target of the hate speech and the text identified in the tweets motivating the label choice. Tweets are either collected randomly from Twitter based on lexicons or reused from

a dataset introduced by Mathew et al. [2019]. Before starting the annotation procedure, the annotators underwent a pilot annotation where they are provided with detailed guidelines and examples. After the pilot, around one-third of the original annotators are chosen for the primary annotation round. For the hate speech task, only hate speech and not offensive language is considered.

The **Online Harassment Corpus** by Golbeck et al. [2017] provides 20,427 tweets annotated for harassment. For the authors, harassment includes threats, hate speech, and direct harassment (language meant to violate a certain group or person directly). To collect tweets, the authors searched for tweets based on a set of terms deemed to be connected to harassment frequently. Every tweet needed to be annotated by at least two human annotators instructed by an extensive codebook, including examples. Since the authors set the threshold comparably high and explicitly excluded offensive language, all tweets labeled harassment are also included as hate speech for the hate speech task.

The **RedditBias** Dataset [Barikeri et al., 2021] is also used in the hate speech task.

3.2.4.5 Gender Bias

Voigt et al. [2018] introduced the **RTGender** dataset, which is a collection of five datasets with comments from Facebook, Reddit, TED, and Fitocracy to study the perception of gender. The five datasets contain unlabeled posts from public figures and their responses, as well as comment response pairs where the gender of the author is known. Voigt et al. [2018] then labeled a small subset of those five datasets using crowd-sourcing. The resulting labeled dataset consists of 15,351 manually annotated posts and responses. For the data collection, only labeled posts and comments are taken and labeled either neutral or biased (if they are annotated with a "positive" or "negative" gender perception).

The **Workplace Sexism** dataset constructed by Grosz and Conde-Cespedes [2020] consists of 1,136 sentences labeled with regard to sexism. 55% of the corpus is a subset of a dataset used by Waseem and Hovy [2016] (also used in the racial bias task), which is labeled for sexism. The rest consists of 25% manually filtered work-related quotes as well as 20% miscellaneous press quotes. This dataset is different from other datasets in that it tries to exclude "hostile sexism," which can be more often found on social media and less in the workplace where the occurring sexism is often more subtle. As in news coverage, the present sexism will also likely be more subtle it is a good fit for a transfer of results toward media bias.

The **"Call me sexist, but..."** dataset [Samory et al., 2020] contains 13,634 sentences labeled on sexism using crowd-workers. This dataset also partly contains

the dataset by Waseem and Hovy [2016]. To avoid duplicates, these are excluded here from the gender bias task collection. The remaining sentences originate from tweets that are filtered based on their beginning with “call me sexist, but...” [Samory et al., 2020]. These tweets are then annotated by MTurk crowdworkers. Five annotators labeled any statement on multiple sexism scales, with a majority needed for a statement to be labeled sexist. The di□erent sexism scales provided are summarized by labeling a sentence gender biased if one or more of the kinds of sexism are present.

Finally, the part of the **RedditBias** dataset on gender is also included in this task.

3.2.4.6 Racial Bias

The race part of the **RedditBias** dataset is included in the racial bias task.

The dataset provided by Waseem and Hovy [2016] is a collection of annotated tweets for hate speech and racism. The dataset is created by collecting tweets that contain specific keywords. The authors then annotated the collected tweets. As only the tweet IDs are published, the actual tweets needed to be accessed via the Twitter API. From the original 16,914 tweets, however, only 7,700 tweets could be retrieved. Since many hateful and racist tweets are deleted from Twitter, only a low share of racist tweets is retrieved. The dataset is, however, still included since the data foundation for the Racial Bias task remains scarce. After concluding the experiments, the Waseem and Hovy [2016] will be removed from the Racial Bias task because it seemingly leads to problems for overall classification results. Therefore, an alternative dataset to extend the racial bias task is of great importance.

The dataset from Ghoshal [2018] is a collection of tweets containing racial bias, which includes tweets from 2018 that have been labeled as containing “racial bias” or “not racial bias” by human annotators. The dataset contains 751 tweets along with their labels and additional information on the user, the location, retweets, and likes. Though it is the only dataset with racially biased tweets publicly available, it lacks clear documentation and a description in a published research paper.

3.2.4.7 Fake News

As the **Liar Dataset** by Wang [2017], which is already described under the cognitive bias task, contains statements labeled on their truthfulness, it is also included under the fake news task.

The **Pheme** dataset [Zubiaga et al., 2017] from the Pheme challenge 2018 and extended by Kochkina et al. [2018] provides a dataset of 5,222 tweets labeled on

the veracity and rumor detection. The Pheme challenge is organized by the Pheme project (led by the University of She□eld) and aims to develop models for analyzing online misinformation. The challenges usually include manually annotated social media posts, and the participants aim at automatically classifying them as rumors or checking their veracity. For the usage in the fake news task, only the veracity of the statements is considered. The extended Pheme version by Kochkina et al. [2018] focuses on nine news events. All annotations are done by journalists who fact-checked all tweets and classified them as false if there is no confirming evidence found [Zubiaga et al., 2017].

FakeNewsNet by Shu et al. [2020] is a collection of articles classified as fake news. Additionally to the truth value of the articles, the authors provide "news content, social context, and spatiotemporal information" [Shu et al., 2020, p.1]. The authors collected articles from multiple fact-checking websites to gather the articles and truth values. Only the articles and their veracity label are used for the fake news task. The articles are split up to a sentence level, resulting in 6,337 sentences.

3.2.4.8 Political Bias

For political bias, the **BigNews** corpus [Liu et al., 2022] and the **UsVsThem** dataset [Huguet Cabot et al., 2021] are used as they both measure political leaning.

Additionally, the **SemEval** dataset introduced by Kiesel et al. [2018] for the SemEval 2019 Task 4, which is focused on detecting partisan news, is added to the political bias task. Partisan news is news coverage that is slanted toward a political ideology. Here, a news article is called partisan if it is either left or right leaning. If an article is annotated as classified, it is considered politically biased for the political bias task. The International Workshop on Semantic Evaluation (SemEval) organizes annual tasks to evaluate computational systems doing semantic analyses. Tasks include, among others, sentiment analysis, semantic role labeling, and, as seen here, partisan news detection. Of the datasets given in 2019 Task 4, only the collection of 645 manually annotated articles is included in the political bias task. Also, available distantly labeled articles from allsides.com are excluded since, with BigNews, a considerable corpus of the same approach is already included. Splitting up the articles resulted in 6,337 labeled sentences.

3.2.5 Data Preprocessing and Balancing

As seen in the previous description of the datasets, the selected datasets varied substantially. Before using them as part of MBIB, a large amount of preprocessing is necessary. For some datasets, only IDs are given, and the tweets or articles needed to be scraped (for example, Waseem and Hovy [2016] and FakeNewsNet). Others

needed to be recombined, sorted, or filtered for relevant articles. All datasets are brought into a uniform shape to facilitate the usage of the datasets. That shape consists of a unique ID of a statement, an ID indicating to which dataset the statement belongs, the text, a binary label, and, if given, additional labels (in a unified form, however, differing in meaning depending on the original label). The biggest preprocessing decision needs to be made when creating binary labels. For some datasets, binary labels are already given (e.g., Spinde et al. [2021c]). A threshold is determined for other datasets with continuous labels to binarize the data. If possible, the author's recommendation for a threshold is followed. Finally, multi-categorical labels are collapsed into two categories (most prominently "right" and "left" into "biased"). Binarizing results into all datasets having the same shape and "biased" vs. "non-biased" labels. The decision to put all labels into a binary format is based on two reasons: (1) It allows an easy combination of different datasets without requiring different model heads. (2) It follows in line with the task principles set up by Wang et al. [2019a] to formulate the task as simply as possible.
The text is preprocessed only rudimentary. Hashtags and tags are eliminated from social media datasets to prevent them from influencing the classification and since news articles usually do not contain hashtags. Furthermore, non-text objects like smilies are filtered out to reduce the overall token size.

The final step in constructing MBIB consisted of balancing the data into a 50:50 relationship of biased vs. non-biased labels. The balancing is done by randomly drawing an equal amount of data points from both categories. The amount sampled is determined by the smaller available task (ensuring that the dataset would be as big as possible). This balancing results in smaller datasets used in the experiments than the entire available data collection. However, the labels are moderately balanced for all tasks, resulting in only a slight data loss. Research suggests that class imbalance can have an impact on the classification accuracy [Li et al., 2010, Padurariu and Breaban, 2019].
Finally, datasets are limited to a maximum size of 500,000 data points. The limit is mainly introduced to reduce the computational effort necessary when training and testing the models. The reduction only applied to the BigNews Corpus and Kaggle Jigsaw datasets. While this might go along with some loss of information, this loss should be limited as these datasets remain the largest.

When selecting the datasets for the tasks, one concern is the differences in dataset sizes per task. Downsampling to the size of the smallest dataset would mean losing large shares of available training and testing data. Not only would this severely reduce the models' capabilities on the downsampled datasets as much potential information on the bias would have been lost, but it would have also impacted the robustness of the results [Soekhoe et al., 2016]. Furthermore, for transformer models, bigger training corpora usually lead to better model performances [Brown et al., 2020, Fortuna and Nunes, 2018].

Alternatively, upsampling techniques are considered to increase the amount of data by, e.g., oversampling smaller datasets. Upsampling, however, would have required enormous repetitions of some small datasets, potentially leading to overfitting. Upsampling would have also required adjustments to the k-fold-cross validation to avoid data leakages, making the experimental setup less comprehensible and harder to reconstruct. That is why neither down- nor upsampling techniques are used. However, whether the dataset size plays a role in the performance will be discussed in section 6.4. After constructing media bias tasks, selecting suitable datasets for each task, and preprocessing them into a unified shape, MBIB's construction is concluded by a framework defining how to evaluate models on MBIB. The framework will be discussed in section 5.1. The following sections will concentrate on choosing feasible models (chapter 4) and setting model baselines on MBIB (chapter 5).

CHAPTER 4
Model Selection

Setting a model baseline on the MBIB tasks shows the current state-of-the-art in media bias detection. Furthermore, it can generate insights into MBIB by seeing how well datasets fit together, whether unexpected results occur and what the best performance measurement would be.

As mentioned above the model baselines here focus only on transformer models. A complete model baseline would include training and testing all available transformer models on the MBIB tasks. However, due to the abundant availability of different models and computational restrictions, five models are selected, fine-tuned, and tested on each task. When choosing models, the standard procedure is often to rely on other researchers' model choices or to argue for a certain model theoretically. Such arguments are also possible here. For instance, auto-encoding models are usually preferred over autoregressive models for classification tasks. Theoretical reasoning would, however, remain speculative. That is why a proxy task is implemented to make an informed model choice based on empirical support.

In the first step, an overview of existing models is created. The transformer taxonomy provided by Kalyan et al. [2021] laid the foundation for the model overview and categorizing transformer models. Additionally, recently published models are added to the overview. Generally, pretrained models are considered as only fine-tuning would be feasible in the experiments. Some of the models are excluded based on specific criteria: (1) The models need to have an implementation available on Huggingface. An available implementation simplifies access to the pretrained models. Also, it ensures that the models are, to a certain degree, widely used and that the results are relevant for future research that wants to base its model choice on them.

(2) Models that are domain-specific in a non-related domain (such as programming or biomedical) are excluded. (3) Also excluded are multi-lingual models for translation tasks and non-English models. (4) The models are always considered in their "base" form. This implies that enlarged and compact models (distilled or pruned models) are excluded. Only base models are chosen to ensure comparability between models and prevent the parameter size from being the decisive performance driver. After applying these criteria, 30 transformer models remained. All of these 30 models are tested in the proxy task.

4.1 Transformer Models

Before discussing the proxy task and its results a theoretical foundation of the structure and architecture of transformer models is necessary. This will enable interpreting the selected models and their performances.

The transformer model by Vaswani et al. [2017] introduces the idea of only using attention mechanisms as the main method to capture linguistic features for language understanding and generation. Unlike in previous models (such as LSTMs or RNNs), where the text is inputted only sequentially, transformers allow for a simultaneous input. An illustration of the original transformer model architecture by Vaswani et al. (2017) can be found in Figure 2. The model consists of an encoder (depicted on the left side of Figure 2) and a decoder part (depicted on the right side of Figure 2). The input sentence is fed simultaneously into the encoder, which passes a vector representation into the decoder. The decoder outputs a prediction (for example, in translation, the first word of a translated sentence). The prediction is fed into the decoder to derive the next prediction (e.g., the second word of the translation). This process is repeated until an end-of-sequence token is generated. This architecture allows for generative learning tasks such as translation or abstract summarization.

The text input into the encoder is first transformed into embeddings. A positional encoding is added to the embeddings, allowing the model to reconstruct better where individual words are placed in the input. The embeddings are then passed into a multi-head attention mechanism. The attention mechanism is supposed to capture linguistic features by weighing the relationships between the inputted embedding and how important an embedding is for that feature. The attention mechanism outputs one attention vector for every inputted embedding. The embeddings are fed into a neural network that outputs a value, key, and query vector for every input embedding. Figure 3 depicts how one attention vector is calculated from the three vectors. This process is done simultaneously for all inputted embeddings. To capture more linguistic features, multi-head attention is used, which repeats the same process multiple times in parallel, each time with independent neural networks to produce other query, key, and value vectors. After several attention vectors are calculated, they are concatenated and linearly transformed before being passed into a

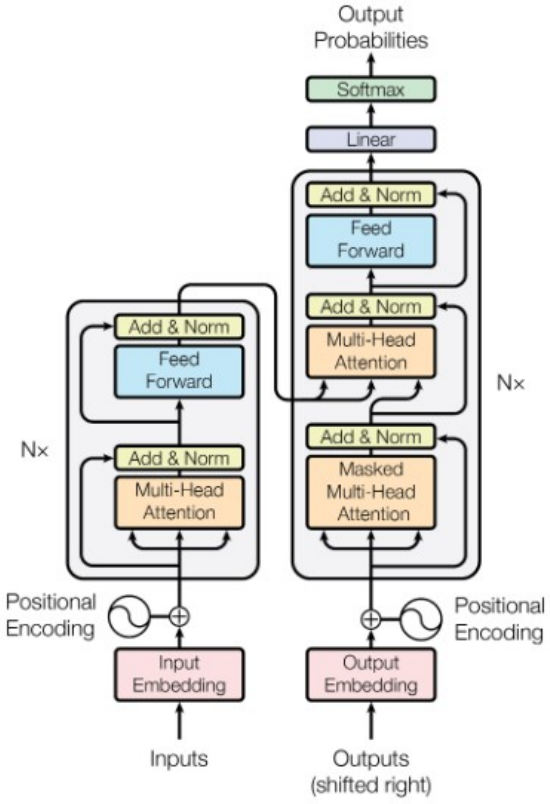


Figure 2. Transformer architecture by Vaswani et al. [2017]

feed-forward neural network. The network's output constitutes the final embeddings produced by the encoder, fed into the decoder.

The transformer's decoder works similarly to the encoder. However, the multi-

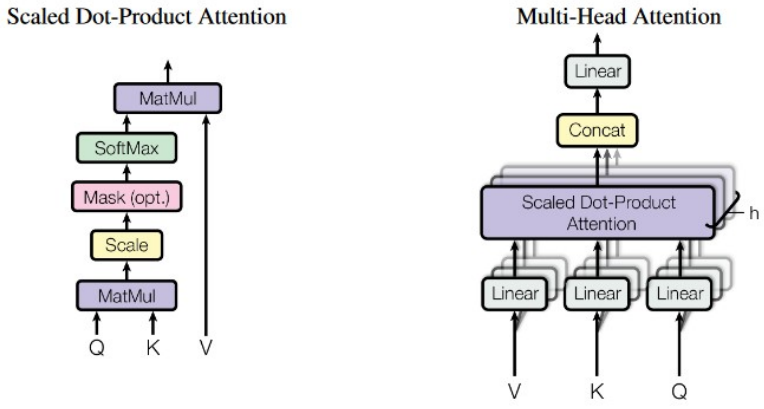


Figure 3. Attention mechanism by Vaswani et al. [2017]

head attention mechanism must be masked when input is passed sequentially. All the embeddings of future tokens (not yet predicted) are replaced with a mask token. The output is combined with the output of the encoder, fed into another multi-head

attention mechanism, and then into another feed-forward neural network. To produce an output, the final embeddings are linearly transformed to the length of the entire vocabulary and passed into a softmax function. The vocabulary item with the highest probability constitutes the model's prediction.
Transformer models are usually trained in two steps. First, the model is pretrained in an unsupervised manner on a large corpus of text using a simple language task objective. Then, the model is fine-tuned using task-specific labeled training data. For fine-tuning, the last layer of the model is adapted for the specific task. The adapted pre-trained model is further trained with the labeled data. This two-step process allows users to adapt models for individual tasks without going through a pervasive pre-training task.

4.1.1 BERT

In the following section, BERT [Devlin et al., 2019], based on the encoder part of the transformer model, is discussed in more detail. Though BERT is not used in the experiments, almost all available transformer models are derivations of BERT or relate to the model in some way. Understanding BERT's architecture and usage are helpful when considering other transformer models. How the models of this survey differ from BERT will be discussed in detail in section 4.4.
Before words are transformed into vector embeddings, they are split up into WordPieces (e.g. "computer" would be split up into "comp" and "##uter"). Splitting up the words allows the model to better understand grammatical changes in words or word compositions, eliminating the need for preprocessing steps such as stemming or lemmatization. After the WordPiece input is translated to embeddings, a positional embedding is added to each token. The final input to BERT is a list of vectors. For BERT the input length is limited to 512 tokens. To ensure that all inputs have the same length, the input is filled with padding tokens ([PAD]) until a predefined maximum length is reached. The input is truncated if the input is longer than the maximum length.
BERT uses the same architecture as the original transformer model's encoder. However, 12 encoders are stacked (each encoder except the first taking as input the output of the previous encoder). The outputs of the last encoder are BERT's final encodings.

As mentioned above, the pre-training of transformer models is self-supervised on unlabeled text data. The corpora used for training vary. BERT is trained on a collection of books and Wikipedia [Devlin et al., 2019]. The attention mechanisms are deterministic, meaning only the weights in the feed-forward networks and those producing queries, keys and values are changed during backpropagation. BERT's pre-training objective consists of masked token prediction and next-sentence pre-

diction. In masked token prediction, some of the tokens in the input sequence are replaced by a mask token. The model's objective is then to predict the masked tokens. The final embeddings are linearly transformed and inputted into a softmax function to get a prediction from BERT. The token with the highest probability is BERT's prediction. The prediction can be compared to the actual masked token, and the error is back-propagated through the model. For the next sentence prediction task, BERT is inputted with two sentences separated by a separator ([SEP]) token. The model should then predict whether the second sentence follows the first or not. A classification token ([CLS]) is added for classification tasks such as next-sentence prediction. The input for BERT then looks like this:

$$[E_{CLS}][E_{This}][E_{is}][E_{mask}][E_{sen}][E_{\#\#tence}][E_{SEP}][E_{This}][E_{mask}][E_{follows}][E_{SEP}]$$

The [CLS] token can, for instance, also be used for fine-tuning a model on media bias classification tasks. In fine-tuning, a classifier layer is put on top of the CLS token, and the pretrained weights are used. Training is then continued on the labeled training data.

4.2 Proxy Task

The proxy task, aimed at filtering model candidates to be evaluated on MBIB, consists of fine-tuning the 30 identified transformer models on the BABE dataset by Spinde et al. [2021b] and comparing their classification performance. As described in subsubsection 3.2.4.1 the BABE dataset has previously been used to measure the performances of transformer models. While less extensive and multi-faceted than MBIB, the dataset is relatively small, requiring little resource and enabling fast training. Also, it is one of the best media bias datasets available regarding annotation quality and annotator expertise. The proxy task resembles the experiment in Spinde et al. [2021b]. All fine-tuning is performed on an NVIDIA A100-SXM4-40GB GPU using 5-fold-cross-validation. In k-fold-cross-validation, the data is divided into k splits of equal size. The model is then trained k-times, each time with a different split as the test dataset, while the other four splits serve as training data. This method ensures, especially for small datasets, that outliers in the data do not influence the performance results. For the proxy task, 30 models are fine-tuned five times resulting in 150 model training and testing. The model training parameters can be found in Table 2. These are also chosen to resemble the training parameters of Spinde et al. [2021b]. The number of epochs is determined by using an early stopping criterion. When the validation loss after an epoch rose compared to the validation loss of the previous epoch, training is stopped to prevent overfitting. A more detailed discussion of the early stopping criterion will be given in subsection 5.1.3. The batch size is usually set to eight. However, it had to be lowered for some large models due to memory constraints.

Table 2. Hyperparameters of the proxy task

Optimizer	AdamW
Learning Rate	$3e^{-5}$
Epochs	Early Stopping
Dropout	10%
Max Length	512

4.3 Proxy Task Results

The training usually only takes two to three epochs until the early stopping criterion is triggered. The F_1-Scores of individual folds are averaged to retain one performance score. The results of the top five performing models can be found in Table 3. In the appendix Table 10 the results of all 30 tested models can be found. The models' performances are close to the performances found by Spinde et al. [2021b]. The top 5 results of the proxy task are used to choose the models which would be fine-tuned and compared on all eight tasks of MBIB. The types of models chosen by the proxy task are diverse. With BART, an encoder-decoder-transformer performs best. With ELECTRA and ConvBERT, there are autoencoding models included. With GPT-2, there is an autoregressive model in the top five. RoBERTa-Twitter, as a social media domain-specific model, also performs well. As some of the datasets of MBIB are based on Twitter data, it will be interesting to see how a model specifically designed for this domain will perform on them compared to the other models.

Table 3. Top-5 performing models on the proxy task

Rank	Model	Average F_1-Score
1	BART	0.81146
2	RoBERTa-Twitter	0.80851
3	ELECTRA	0.80646
4	GPT-2	0.80371
5	ConvBERT	0.80321

4.4 Models

The following section describes the five transformer models chosen by the proxy task. The main focus lies on how their architecture and pretraining objectives differ. An overview of the size and pretraining corpora used for each model can be found in

Table 4.

Table 4. Parameter overview of the top-5 models

Model	Parameters	Vocabulary Size	Pretraining Corpus
BART	140M	50,265	Books (2015), Wikipedia
RoBERTa-Twitter	125M	30,522	Books (2015), CC-Stories (2019), CC-News(09.2016-02.2019), Open Web Text (2018), Wikipedia, Twitter
ELECTRA	110M	30,522	Books (2015), CCa, ClueWeb 2012-Bb, Wikipedia, Gigaword 5 (2012)
GPT-2	1.5B	50,257	OpenWebText (2018)
ConvBERT	96M	30,522	OpenWebText (2019)

4.4.1 BART

BART is introduced by Lewis et al. [2020] and is a sequence-to-sequence transformer model. This means that BART combines a bidirectional encoder and autoregressive decoder, like the original transformer model introduced by Vaswani et al. [2017]. BART stacks six encoder and decoder layers in its base model. BART is pretrained using corrupted documents and back-propagating the cross-entropy between the reconstructed document (the decoder's output) and the original document. Lewis et al. [2020] propose multiple corruption methods such as masked tokens, token deletions, or sentence permutations. In comparing the pretraining techniques, the authors find that a permutation of the input sentences and text in-filling where spans of text are masked are the most e□ective corruption methods.
BART can be fine-tuned for numerous downstream tasks ranging from text translation to sequence generation and classification. For sequence classification, "the same input is fed into the encoder and decoder, and the final hidden state of the final decoder token is fed into [a] new multi-class linear classifier." [Lewis et al., 2020]. For classification, like in BERT, a classification token is added at the end of the sequence so that the entire input can be considered in the decoder.

4.4.2 RoBERTa-Twitter

The RoBERTa-Twitter model introduced by Barbieri et al. [2020] is based on the RoBERTa transformer model introduced by Liu et al. [2019], which is trained on 60 million tweets. RoBERTa is an advancement of the BERT model which mainly introduces four changes to BERT's pre-training: "(1) training the model longer, with bigger batches, over more data; (2) removing the next sentence prediction objective;

(3) training on longer sequences; and (4) dynamically changing the masking pattern applied to the training data" [Liu et al., 2019].
The first and third change is based on the finding that increasing data input and training epochs improves the model's performance. A result that has often dominated model enhancement over the last years (compare Goetze and Abramson [2021]). The second change gets rid of the next sentence prediction during fine-tuning. Liu et al. [2019] show that the BERT model is equivalently good when only focusing on masked token prediction. The last change is the introduction of dynamic masking. In BERT, a static mask is assigned during preprocessing. To avoid duplicate masking while training, the mask is randomly reassigned for every input in dynamic masking.
To train, RoBERTa-Twitter Barbieri et al. [2020] take the pretrained RoBERTa base model and retrain it on tweets. They compare this approach to the RoBERTa baseline and a RoBERTa model solely trained on tweets. The final RoBERTa-Twitter model outperforms these approaches on several tasks, such as hate detection, sentiment analysis, and emotion classification. Since these tasks are similar to those of media bias and a lot of the collected data stems from Twitter, the proxy task confirms the potential of this model for media bias detection.

4.4.3 ELECTRA

ELECTRA, developed by Clark et al. [2020], is a transformer model based on contrastive learning. Like BERT, ELECTRA is based only on the encoder part of the original transformer model [Vaswani et al., 2017]. First, a generator model is used for pretraining that replaces tokens of an input sequence. Then a contrastive model is trained, tasked to detect which tokens are replaced by the generative model. Using the generator model ensures that the replaced tokens are closely related to the original input. Clark et al. [2020] claim that this approach allows the model to learn from all the tokens of the input sequence, while previous models based on masked language detection only learned from the few masked tokens.
The generator model is still inputted with a masked sentence and predicts the masked tokens. This prediction is used to train the generator and as input for the discriminator. The discriminator then predicts whether the input tokens are original inputs or are generated by the discriminator. After pretraining, only the discriminator is fine-tuned for downstream tasks. Training two models also means twice the computational cost. The generative model is kept smaller than the contrastive model to reduce the pretraining capacities needed. Both models share the same token embeddings as a weight-sharing strategy, with both models having different sizes.
The authors find that the pretraining outperforms other approaches, especially if the goal is to train a small model.

4.4.4 GPT-2

Of the five models, the GPT-2 model is the only purely autoregressive model that is based on Vaswani et al. [2017] 's transformer model's decoder. It is introduced by Radford et al. [2019] and a further development of the GPT model by Radford et al. [2018]. Based on the transformer's decoder GPT, the input is fed into GPT sequentially into masked self-attention blocks. The pretraining objective of GPT is causal language modeling (which, in this case, means predicting the next token of the input). The masking of the self-attention blocks prevents the model from 'seeing' information on future tokens that are not fed as input yet. After pretraining, the GPT model can be fine-tuned for a downstream task like the other transformer models. A linear layer is added to the model's output embeddings for classification. The main difference between GPT-2 compared to GPT is a substantial increase in model size, which makes GPT-2 the largest model concerning parameters of the five models. Furthermore, Radford et al. [2019] created a new training corpus, OpenWebText, to diversify the training data better.[1]

4.4.5 ConvBERT

ConvBERT [Jiang et al., 2020] is an autoencoder model based on BERT by Devlin et al. [2019]. Its main difference from BERT lies in partly replacing self-attention heads with convolution-based heads. The approach is grounded on the finding that a large share of attention heads in BERT learn redundancies when extracting local features [Jiang et al., 2020]. This is partly due to the self-attention mechanism calculating dependencies between all input tokens, even though only negligible relationships exist for a large share of tokens. In the convolution approach, the softmax result and the value vector of the attention mechanism are not multiplied but passed into a convolution function, reducing its dimensionality and overall parameter size. However, a simple convolution approach would mean that "kernel parameters would be fixed for any input token, not favorable for capturing the diversity of the input tokens" [Jiang et al., 2020]. The authors, therefore, introduce a dynamic convolution method that generates a new convolution kernel for every input token.

[1] GPT-2 is likely the most surprising finding of the proxy task, where usually autoencoding models are the 'go-to' model choice. One explanation might be found in the substantially different model size between GPT-2 and the other models. Like Goetze and Abramson [2021] show, bigger models usually lead to better results. It, however, also shows that only dismissing models on theoretical grounds may fall short. The results of the experiments on MBIB later show that GPT-2 underperforms when tested on MBIB. While this might call into question the generalizability of the proxy task, the general similarity of the results suggests that the proxy task might be less important than originally assumed. If the model choice does not significantly influence the performance, then it also matters less which models are used to set the performance baseline.

To keep the ability to capture global dependencies in the final model design, the authors use a combination of traditional self-attention heads and the new convolution mechanism, passing the input into both mechanisms and concatenating the output. They find that ConvBERT performs comparable or better than ELECTRA and BERT on various standard language tasks while having fewer parameters. ConvBERT with its similar architecture to BERT is a close choice while expecting advantages in lower memory requirements and faster training times for classification tasks such as media bias classification.

CHAPTER 5
Experimental Design

The experiment aims to set a model baseline on MBIB, the media bias benchmark created. It does so by evaluating the five candidate models on all eight media bias tasks of MBIB. The experiment brings together the MBIB tasks created in chapter 3 and the models selected in the proxy task described in chapter 4. The best-performing model on each task will set the model performance baseline. The experiments should also bring insights into the composition and properties of MBIB and the models. For this, the experiment should answer four questions: (1) Is there an overall best model for media bias detection? (2) How do models compare on other metrics besides performance? (3) How much do individual datasets influence the overall score of a task? (4) How important is the size of a dataset for the models' performance on it?
Questions (1) and (2) aim at giving researchers who need to choose a model for a media bias task a foundation to make an informed model choice. Questions (3) and (4) aim to investigate how the combination of different datasets affects the benchmark as this combination is a key difference of MBIB to comparable benchmarks such as GLUE [Wang et al., 2019b].
An evaluation framework is needed to compare models on MBIB that defines how the model should be fine-tuned and which metrics should be used and reported. The framework introduced here uses stratified k-fold-cross-validation on the preprocessed and balanced MBIB data. Overall F_1-Scores, training time and memory usage details, and dataset-specific predictions that allow a breakdown of the overall results are reported. The evaluation framework is introduced in section 5.1. In the remainder of the section, the hyperparameters used in the model training are presented (subsection 5.1.3 and section 5.2). Finally, measures to reduce memory requirements and accelerate training as well as the overall training environment are described (section 5.3 and section 5.4).

5.1 The Evaluation Framework

5.1.1 Stratified k-fold-cross-validation

The models are trained and tested using stratified 5-fold cross-validation. Cross-validation is solely used for evaluating models [Refaeilzadeh et al., 2016]. In cross-validation, the existing data is split into a training and a validation set. The training set is used here to fine-tune the model. A generalizable performance estimate is obtained by comparing the predictions of the model to the actual labels of the validation set. Cross-validation ensures that the measured performance stays generalizable.[1] A generalizable estimate is essential for comparing the performance of multiple models [Refaeilzadeh et al., 2016].

One common variant of cross-validation is k-fold-cross-validation [Anguita et al., 2012, Refaeilzadeh et al., 2016]. The data is split into k folds of equal size, and each fold is used once as the validation set while the other folds are used as the training data. This process is repeated until every fold has served as a validation set once. Each repetition is independent, meaning the model is trained from scratch each time, so it never trains on the same data it validates.[2] A final performance score is obtained by averaging the scores of each fold. K-fold-cross validation is more robust than standard cross-validation approaches as it ensures that outliers and uneven distributions are included in the validation data at least once. Kohavi [1995] show that a stratified 10-fold-cross validation delivers the least biased model estimates compared to the standard cross-validation approach and classical k-fold-cross validation.

According to [Refaeilzadeh et al., 2016, p.3], "stratification is the process of rearranging the data as to ensure each fold is a good representative of the whole." Here, stratified k-fold-cross-validation is used to ensure that all datasets are represented in every fold. In regular k-fold-cross-validation, the data assigned to each fold is randomly drawn. In stratified k-fold, the data is also drawn randomly, but the class distribution is maintained [Diamantidis et al., 2000]. Stratified k-fold-cross validation is explicitly recommended by Kohavi [1995] when k is smaller than 10. Usually, the data is stratified by the classes (or the labels). Since the labels are binary and balanced, this is not necessary. Stratification is applied to maintain the dataset distribution instead. If one dataset constitutes 10% of the entire data, then each fold will consist of 10% of that dataset. Stratification ensures that small datasets are su□ciently distributed between training and validation datasets. This is important since the di□erent datasets of each task cover di□erent facets of bias. If a validation set did not contain any data points of a small dataset, it would reduce

[1] Measuring only the performance of data used in training would likely yield better results than the model would achieve on unseen data.

[2] Each model is, therefore, fine-tuned k times from scratch.

the generalizability of the results. A k of size five is chosen as a trade-o between a k as big as possible while remaining computationally feasible. A k of size 5 is also common in the literature [Fushiki, 2011].

5.1.2 Metric

Finally, a performance metric needed to be defined for MBIB. The authors of SuperGLUE define for the evaluation that "Tasks must have an automatic performance metric that corresponds well to human judgments of output quality" [Wang et al., 2019a, p.4]. Several well-established metrics exist for the tasks' binary classification objective. One possible metric would be accuracy, which indicates the share of classified statements. However, when classes are unbalanced, the accuracy does not adapt and can lead to misleading scores.[3]
Therefore, F_1-Scores are used as the metric. The F_1-Score takes the harmonic mean of the prediction and recall [Chinchor, 1992]. The precision is the share of data points correctly classified as biased from all classified as biased. Recall measures the share of correctly classified biased data points from all biased data points. The F_1-Score ensures that high amounts of false negatives and false positives play a more visible role in the score. The F_1-Scores is also a well-used metric in SuperGLUE [Wang et al., 2019a].

5.1.3 Early Stopping

Determining the appropriate number of training epochs can be di cult before training. Setting the epochs too low risks that the model does not learn su ciently and misses out on possible further improvements. Setting the number of epochs too high increases the possibility of the model overfitting and thus not generalizing well [Prechelt, 1998]. An indicator for overfitting is the validation loss. The validation loss is the accumulated error of a model's predictions and must be calculated on data independent of the training data. If a model begins to exhibit a rising validation loss, it is indicative of overfitting [Prechelt, 1998]. The models used here are trained in epochs, each consisting of one iteration over the training data. The validation loss is calculated after each epoch on the independent validation set.
Stopping the training after the epochs will prevent overfitting. One way to find this amount is to set a high number of epochs and then determine after how many epochs the loss saturates. One could then retrain the model or restore the weights of that epoch. However, when each epoch takes up a substantial amount of resources and

[3] When for example, a dataset consists of 80% class A, and 20% class B and the model predicts only class A, an accuracy value of 0.8 would mislead to think the model is working correctly. Even though the data is balanced before training, media bias is generally a field where heavily undistributed classes are found. Furthermore, when only looking at the performance on individual datasets, the classes will no longer be balanced.

time, like here, this requires much time and computational expenses. To minimize training time, an automated way to prevent overfitting is used. The validation loss is measured periodically, and the model training is stopped after a validation loss starts to rise. This method is called early stopping [Prechelt, 1998].
The early stopping is complemented with a patience variable: if the validation loss rises in two consecutive epochs, the training is stopped. The patience of two epochs ensures that the training is not stopped when a small stagnation occurs, after which the validation loss drops again [Zhou et al., 2020]. Early stopping does not guarantee that a global minimum of the validation loss is reached, so setting a patience variable increases the confidence in the minimum found. Though early stopping eliminates the need for setting the number of epochs, introducing patience replaces it with another hyperparameter for which there is no clear best value. Setting the patience to two represents a trade-off between models being trained to the best amount of epochs while holding the training within limits.

In classical k-fold-cross-validation, the data is split into folds used as a train and a validation set. However, when using early stopping and calculating the validation loss on the validation set, one uses the validation data for hyperparameter tuning. This might endanger the independence of the validation set. According to [Bishop, 1995, p. 372] when optimizing a model using the validation set, "this procedure can itself lead to some overfitting to the validation set, the performance of the selected network should be confirmed by measuring its performance on a third independent set of data called a test set." Berrar [2019] describes how this split can be transferred to cross-validation by sampling a validation set from either the training data or the left out fold (called validation set previously) and using the remainder of the left out fold as the test set. Since the folds are quite big in this case, the left-out fold is split up into a validation and a test set (at a 1:3 ratio). The validation loss for early stopping is only calculated on the validation set. The final score is only calculated on the fully independent test set. Figure 4 depicts the 5-fold-cross validation architecture used.

5.2 Hyperparameter Choices

The batch size depends on the maximum input length to the model (since longer inputs would require more memory). That is why a text length analysis is conducted for every task. The cut for the maximum input length is set at the 99th percentile of the text length. The input length ranged from 34 tokens (Racial Bias) to 294 tokens (Linguistic Bias). On average, the 99th percentile cut-off is after 144 tokens. The text sources can largely explain the difference. For example, tweets (most common in the Racial Bias task) are usually shorter than article sentences (most prominent in the linguistic bias task). All tasks, however, remained well below 512 tokens, the

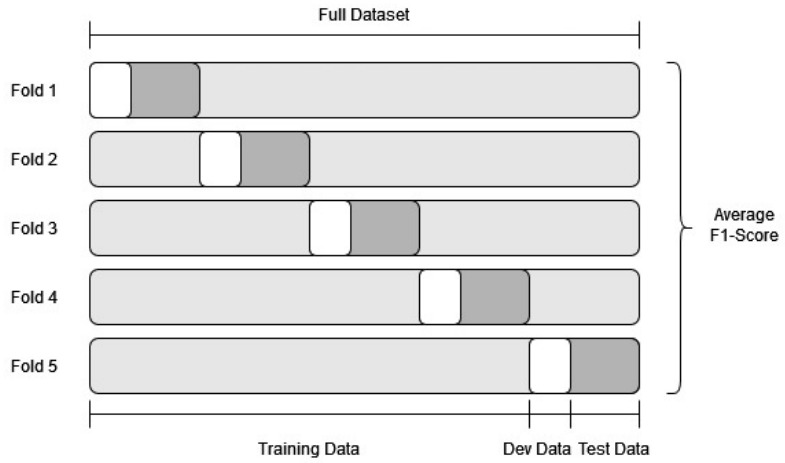


Figure 4. Depiction of the evaluation framework

maximum input length for many models.

The upper limit for choosing a suitable batch size is the memory capacity of the GPU used. The memory needed is correlated to the input length (as also shorter sentences are padded to this length) and the parameter size of the model. The lower limit is set by time constraints, as smaller batch sizes extend training time extensively. In section 5.3 measures to raise the upper limit are described by reducing the overall memory requirement. However, a batch size that is too high can lead to a model not learning properly, often displayed by a loss that does not converge. To find the best trade-o, on the BABE dataset, linearly batch sizes from 512 downwards are tested, and the learning behavior is analyzed using the Weights&Biases interface. A batch size of 64 is found to have the best trade-o between fast training and stable learning behavior for all models.

Similar to the batch size, there is not a singular best rate for the learning rate. When fine-tuning BERT Devlin et al. [2019], use a learning rate of $5e^{-5}$. This rate is also used by Spinde et al. [2021b]. Individual models could be optimized on the learning rate with, e.g., a grid search. However, doing this for all models is computationally expensive and could distort comparisons between models. That is why the learning rate is set to $5e^{-5}$ for all models.
A learning rate scheduler is used to reduce the importance of a fixed learning rate. A learning rate scheduler gradually reduces the learning rate over time. Seong et al. [2018] show that using such a scheduler leads to faster convergence of the loss. Furthermore, it reduces the need for extensive hyperparameter tuning while still resulting in optimal training results [Yedida and Saha, 2019]. Since, due to the early stopping criterion, the exact length of the model training remains unknown, a linear learning rate scheduler from start to finish could not be implemented. The learning rate scheduler is initialized with a cosine function instead. The cosine function lets

the learning rate drop quicker at the beginning and then converge but never reach zero.

A table of the parameters used in model training can be found in Table 5.

Table 5. Hyperparameters evaluation on MBIB

Optimizer	AdamW
Learning Rate	$5e^{-5}$
Epochs	Early Stopping
Dropout	10%
Max Length	Input dependent ($\leqslant 512$ Tokens)
Batch Size	64

5.3 Training acceleration

Since training all models is a significant limiting factor, multiple measures are implemented to reduce memory usage and speed up training. Memory usage is correlated to the training time as, for example, lower memory usage allows for larger batch sizes.
One measure to reduce memory usage is to use gradient checkpointing. In gradient checkpointing, only some activations in the forward pass of the model are saved. Those not saved are then recalculated on the backward pass. Gradient checkpointing reduces memory usage, however, increases the overall calculation effort. The advantage here is that it still improves speed because it allows for larger batch sizes. Furthermore, larger batch sizes can lead to a more stable training. The same trade-off applies to gradient accumulation, where the batch is split up into multiple forward and backward passes while the gradients are accumulated. Only when all gradients are accumulated the optimizer takes another step. Gradient accumulation allows for much bigger batch sizes while slowing down the computation.
Finally, to further reduce memory usage and increase the training speed, Huggingface's Accelerator class is used. It not only allows for the handling of gradient accumulation and checkpointing in PyTorch but also offers mixed precision training (introduced by Micikevicius et al. [2018]) by reducing the memory complexity of activations down to fp-16 (instead of fp-32). The memory optimization increased the batch size by a factor of eight.

5.4 Training Infrastructure

The models are trained and tested on NVIDIA A100-SXM4-40GB GPUs. Every model is trained on a single GPU, though multiple models are trained in parallel on different GPUs. Training time and GPU utilization are measured during training. The pretrained models, as well as their respective tokenizers, are all loaded from Huggingface's API. The models are used in their base variant (as in the proxy task). A list of the exact models used can be found in Table 9.

CHAPTER 6
Results

6.1 Overall Performance

The average F-1 Scores, the variance between folds, and the average training time per fold can be found for all models and tasks in Table 6. The total training time for all models and tasks added up to 114h.

In each task, the classification results lie close to each other. The biggest difference between the best and worst performing models is in the fake news and the political bias tasks, with an F_1-Score difference of ~ 0.08. However, the difference remains within a 0.025 margin for all other tasks. Also, no one model dominates all other models in performance for all tasks. Interesting is the overall performance of all models on the fake news task. Here, the worst performance is observed (average F_1-Score over all models of 0.66). This result correlates with the initially mentioned intuition that more than linguistic features are needed for successful fake news detection. Adding real-world knowledge will likely improve the performance on this task. However, it shows that, to a limited extent, linguistic features contain information about fake news. The models can extract this information and perform better than if they had guessed whether a statement is true.

The second-worst performance is reached by the tasks of political bias (average F_1-Score of 0.69) and cognitive bias (0.70). Both tasks should be interpreted similarly, as they share most of their data. The performance on the political bias task is surprisingly low (especially seeing that the task only consisted of predicting if there is a bias, not whether it is "left" or "right"). However, the results might be explained by the relatively noisy data, especially from the distantly labeled BigNewsCorpus, which comprised a large share of the data. An analysis of each dataset's predictions

separately will be conducted in section 6.3. The linguistic bias task also did not yield high classification results. An intuition on the causes, however, requires further analysis.

The models performed well on the racial bias (average F_1-Score 0.87) and hate speech (0.88) detection. Both tasks exclusively consist of datasets based on scrapes from social media. These datasets collected for hate speech and racial bias contain many racial slurs, swear words, and offensive language. Such bigotry might make it relatively easy for models to identify potentially racist or hateful posts. However, this does pose a problem with the generalizability of media bias. The racism and hate projected in media are likely more subtle and much harder to identify than in social media. While showing that detecting racism and hate works well, this also calls for creating datasets in the media bias domain, specifically tackling these topics.

Table 6. Average F_1-Scores

	Linguistic Bias			Cognitive Bias		
	Model	F_1-Score (Variance)	Avg. Training Time per Fold	Model	F_1-Score (Variance)	Avg. Training Time per Fold
1.	ConvBert	0.7126 (1.15E-06)	1h23m	ConvBert	0.7044 (1.13E-05)	0h58m
2.	ELECTRA	0.7122 (1.37E-05)	1h6m	Bart	0.7042 (1.87E-05)	1h11m
3.	Bart	0.7106 (2.09E-06)	1h46m	Roberta-Twitter	0.7006 (4.44E-06)	0h48m
4.	Roberta-Twitter	0.7102 (2.49E-05)	1h12m	GPT2	0.6976 (6.97E-06)	1h16m
5.	GPT2	0.7011 (1.99E-06)	2h7m	ELECTRA	0.6777 (9.14E-06)	0h14m

	Text Level Context Bias			Hate Speech		
	Model	F_1-Score (Variance)	Avg. Training Time per Fold	Model	F_1-Score (Variance)	Avg. Training Time per Fold
1.	ConvBert	0.7697 (1.28E-03)	0h7m	Roberta-Twitter	0.8897 (8.72E-07)	0h37m
2.	Roberta-Twitter	0.7689 (1.03E-03)	0h5m	GPT2	0.8824 (3.72E-07)	1h12m
3.	Bart	0.7622 (6.85E-05)	0h7m	ELECTRA	0.8821 (1.05E-06)	0h40m
4.	ELECTRA	0.7532 (1.33E-03)	0h5m	ConvBert	0.8805 (4.45E-06)	0h58m
5.	GPT2	0.7447 (1.44E-03)	0h10m	Bart	0.8797 (3.61E-06)	1h16m

	Gender Bias			Racial Bias		
	Model	F_1-Score (Variance)	Avg. Training Time per Fold	Model	F_1-Score (Variance)	Avg. Training Time per Fold
1.	Roberta-Twitter	0.8334 (3.73E-05)	0h3m	ConvBert	0.8772 (3.12E-05)	0h1m
2.	Bart	0.8333 (1.14E-05)	0h4m	ELECTRA	0.8768 (5.41E-05)	0h1m
3.	ELECTRA	0.8305 (3.00E-05)	0h3m	Roberta-Twitter	0.8728 (7.48E-05)	0h1m
4.	ConvBert	0.8257 (1.17E-05)	0h4m	Bart	0.8693 (1.08E-05)	0h2m
5.	GPT2	0.8134 (2.04E-05)	0h4m	GPT2	0.8508 (1.37E-04)	0h2m

	Fake News			Political Bias		
	Model	F_1-Score (Variance)	Avg. Training Time per Fold	Model	F_1-Score (Variance)	Avg. Training Time per Fold
1.	Bart	0.6811 (1.77E-04)	0h2m	ConvBert	0.7041 (2.32E-06)	1h15m
2.	ConvBert	0.6787 (1.27E-04)	0h1m	Roberta-Twitter	0.7021 (4.87E-06)	0h52m
3.	Roberta-Twitter	0.6721 (1.84E-04)	0h1m	Bart	0.6997 (5.30E-06)	0h55m
4.	ELECTRA	0.6574 (6.18E-04)	0h2m	GPT2	0.696 (4.55E-06)	1h21m
5.	GPT2	0.6094 (2.27E-04)	0h2m	ELECTRA	0.6255 (2.12E-02)	1h2m

Finally, on the gender bias (average F_1-Score 0.83) and text-level context bias (0.76) tasks, models show a mediocre performance. Both tasks consist of a mix of news and social media datasets. While seemingly not as easily identifiable as racial bias and hate speech, models perform better on them than on the lowest four tasks. The variance remains low for all models and tasks. An indicator that there is only a minimal difference in performance between folds. The size of the training data mainly determines the training time. Still, there are apparent differences between models. The difference shows that the autoregressive models (GPT-2 and BART) require longer training.
Apart from the size of the data the training time is determined by how quickly the validation loss saturates and the early stopping criterion is invoked. Table 7 shows the average number of epochs for each model and task, after which the early stopping is invoked, and the training stopped. Mostly, it took only 2-4 epochs for the loss to saturate. GPT-2 needed the most epochs for all tasks, followed by BART. Therefore, their longer training times can be explained by needing more training epochs.

Table 7. Average number of training epochs

	Linguistic	Cognitive	Text-Level Context	Hate Speech	Gender	Racial	Fake News	Political
Bart	2.6	2.2	3.2	3.4	3.6	3.6	4.2	3.2
Roberta-Twitter	2.4	3	2.8	2.2	2.6	3.2	3	3
ELECTRA	2.2	2.8	3	2.4	3	3	2.8	3.2
GPT2	3.2	3.6	4.6	3.4	2.8	5.4	4.2	3.6
ConvBert	2	2.4	3	2.6	2.8	3.2	2.8	2.4

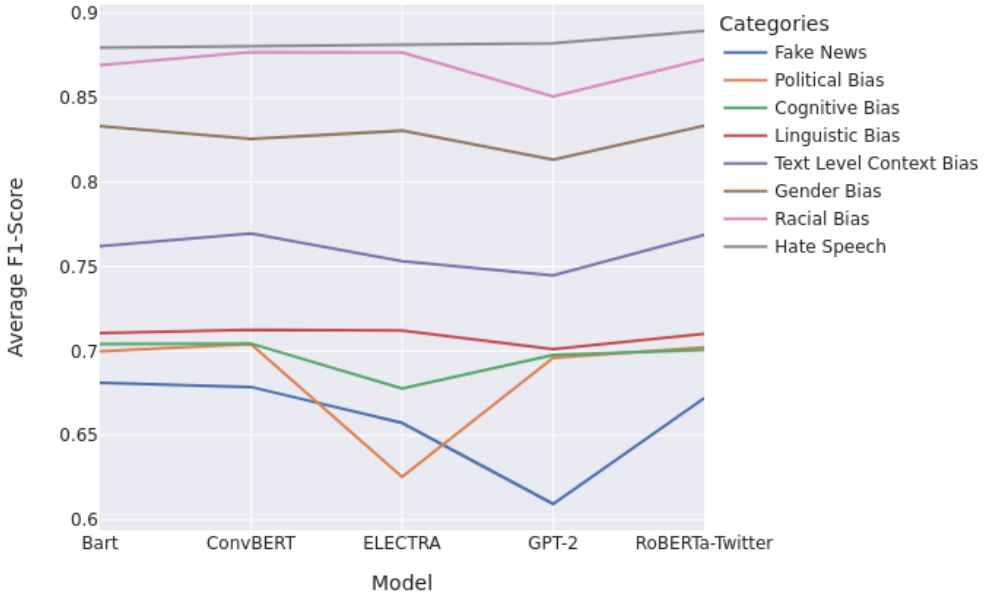
6.2 The Best Model

When looking at the performance results, no model outperforms all others in all tasks. But is there a model more suitable for media bias detection than the others?

What seems immediately clear is that GPT-2 performs worst on five out of eight tasks except on hate speech and not well on the other tasks. This observation confirms the prevailing opinion of the literature that autoregressive models are not competitive on classification tasks. Though there is a clear difference, this is small.

The best-performing model on all tasks is not as easily identifiable. ConvBERT performs best in five out of eight tasks, comparatively low, however, on hate speech and gender bias. Roberta-Twitter performs best on those two tasks but only lies in the midfield for the rest. Figure 5 visualizes the results from Table 6. It confirms that GPT-2 seems to underperform the other models. Furthermore, ELECTRA seems to have a clear performance drop on political and cognitive bias (two tasks that share data). All in all, which model one chooses seems to only slightly impact performance.

Figure 5. Average F_1-Score per model



6.3 Per Dataset Analysis

No big differences between models are detected within tasks. Since each task consists of multiple datasets, it might give a fruitful analysis of how well the models did on each dataset separately. Such an analysis gives insights into whether the performance is similar on the datasets of a task or whether one dataset heavily influences the final result. Furthermore, it can show whether the different size of datasets poses a problem for the benchmark. If there is a link between performance and size of the dataset, then one would expect better performance on bigger datasets.

For this analysis, in each fold, all predictions and actual test values are saved together with their dataset ID. This data allows for calculating F_1-Scores for every dataset and task. Since for the F_1-Scores, the predictions of all folds are combined, they are calculated on predictions over the entire dataset. These predictions are made with models trained on the entire data of one task, so they do not represent a single model being only trained and tested on the respective dataset alone. There might even be a negative transfer: A model trained on all datasets might perform worse on the test set of a particular dataset than a model only trained on that dataset. BABE, where there are reference results from the proxy task, shows this is the case to a small degree.

Figure 6 shows boxplots of the calculated dataset F_1-Scores for each model (so

if datasets appeared in multiple tasks, they are combined for this visualization). Interestingly, this somewhat revises the impression of overall performance gained by only looking at the overall scores. BART, GPT-2, and RoBERTa-Twitter have the highest medians, while ELECTRA has the lowest. ConvBERT and RoBERTa-Twitter seem to have the lowest variance in performance between datasets.

Figure 6. Boxplot F_1-Scores per dataset

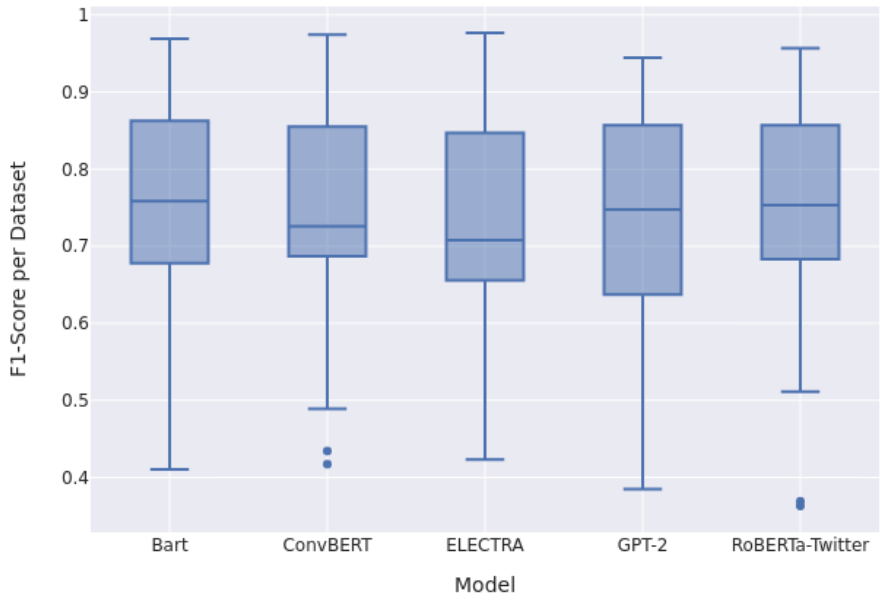
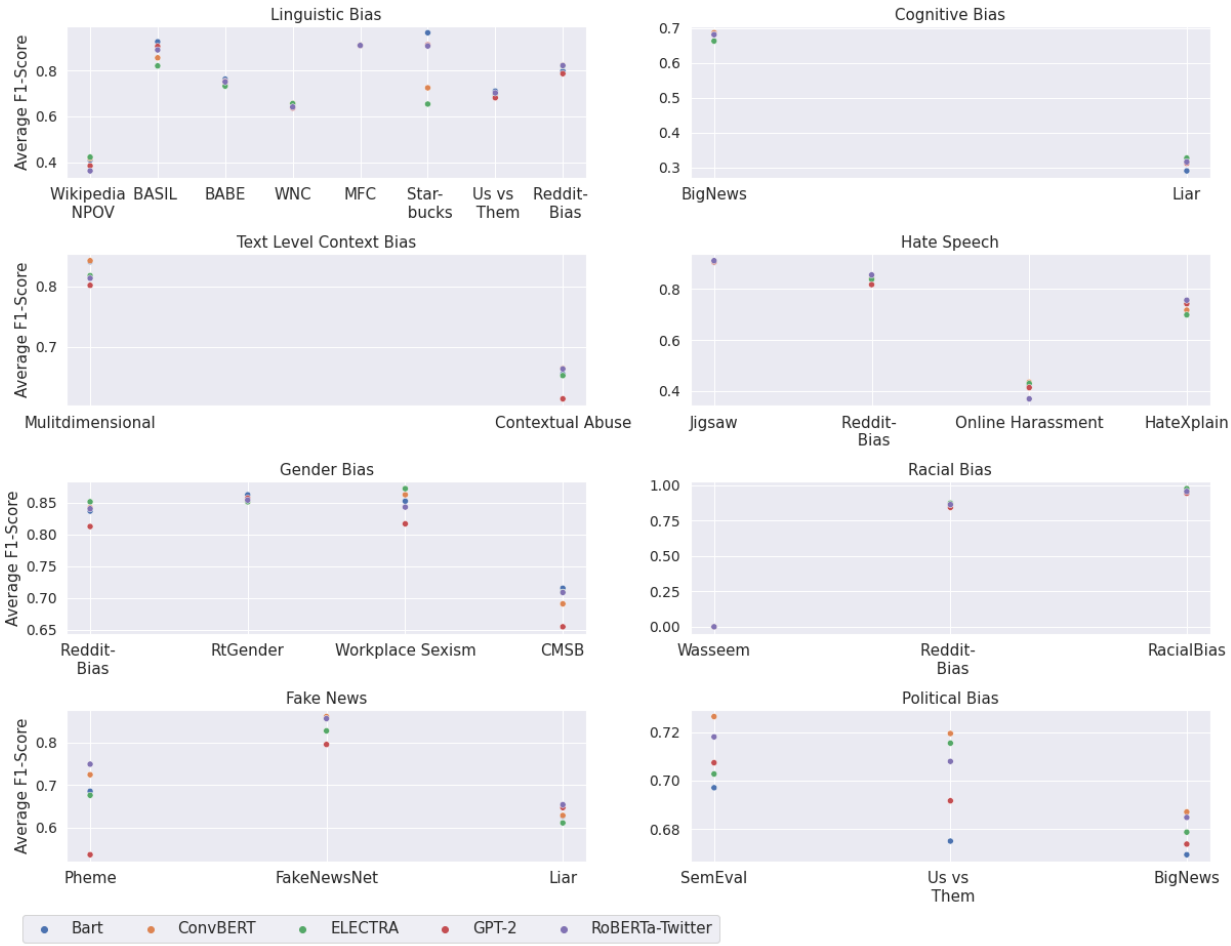


Figure 7 shows the five models' performance on every dataset split up into individual tasks. As opposed to Figure 6, the F_1-Scores are calculated for datasets appearing in multiple tasks separately. Noticeable is that singular datasets show low performance compared to the other datasets of the task. For linguistic bias, this is Wikipedia NPOV. While the Wikipedia NPOV dataset uses the same data source as the Wikipedia Neutrality Corpus (WNC), the models seem to detect substantially less information on bias. The Liar Dataset shows substantially lower results in the cognitive bias task than the BigNews Corpus. Interestingly, the Liar Dataset shows a much higher performance in the fake news task. This performance might be caused by positive transfer during model training. For cognitive bias, the final score is dominated by the bigger BigNews Corpus. This discrepancy reiterates the need for better training data customized toward cognitive bias. The same applies to text-level context bias. Though the di□erence is smaller, the final score is dominated by the bigger Contextual Abuse dataset.

For hate speech, the Online Harassment Corpus shows a comparatively deficient

Figure 7. F_1-Scores per dataset and model



performance. A more detailed analysis of the performance on the Online Harassment Corpus revealed a relatively high accuracy (on average 0.744), however, a low precision (0.519), and recall (0.346). The low recall is due to a high percentage of false negatives. Around two-thirds of actually biased sentences are classified as non-biased (on average, only 11.5% of non-biased sentences are labeled as biased). This high rate of false positives can indicate that the threshold set by the model for what constitutes hate speech is too high for this dataset. The threshold refers to the point at which enough indications of bias are present to label a statement as biased. Annotators and creators of di□erent datasets are likely to set di□erent thresholds determining if they call a statement biased. The CMSB dataset in the gender bias task shows the opposite e□ect: A high rate of false positives leads to a low precision (on average 0.620), while recall (0.794) and accuracy (0.905) are

relatively high. This result might indicate that the models' threshold for classifying gender bias is too low for the dataset. Both differences, for the Online Harassment and CMSB dataset, suggest that the definition applied by the creators as to what constitutes hate speech and gender bias differed from those of the creators of the other datasets.
These findings indicate that the combination of datasets introduces noise into the data through differing bias definitions and standards. One discriminator on top of multiple datasets is then likely going to perform worse than on individual datasets. One response to this could be to discard all but one dataset. As discussed above for neither task a single extensive dataset exists. There is not even a consensus definition of the biases nor thresholds for what should constitute a biased threshold. This is also the reason why datasets differ in the first place. Since unclear boundary cases are going to continue to exist a combination of different definitions and interpretations might average a more balanced understanding than arbitrarily choosing one definition. If a single bias label is required then this approach might bring the classifier closest to the truth.

In the racial bias task, the Wasseem Dataset produced a surprising result: F_1-Scores of 0.0 for all models. Investigations into this revealed fundamental problems with the dataset. Similar to other Twitter datasets, only a tweet ID is given, and all tweets had to be scraped from Twitter by that ID. The scraping returned fewer tweets than originally in the dataset. As the dataset labeled racial bias, it is likely that Twitter internally deleted tweets identifiable as racist. The remaining racist tweets are likely those not violating Twitter's guidelines, resulting in no true positives. The entire dataset should therefore be discarded when using the benchmark. That some positives are left either indicates poor annotation quality or missed removals by Twitter.

Interestingly, in the political bias task, the performance on SemEval, which contains manually annotated news articles, is higher than BigNews, consisting of distantly labeled articles. This demonstrates that the labeling technique makes a data quality difference visible in the classification results.

6.4 Dataset Size and Performance

Since during the combination of datasets for the individual tasks, a concern is the differing size of datasets, an analysis of dataset size and the models' performance on the dataset can give insights into whether the model size is a driving factor for the performance. If there is such a relationship, it will pose a problem for the meaningfulness of the benchmark. Scores of smaller tasks with fewer data could not be compared with those with more data. The score would not be determined by how well a model can learn but by how much training data it gets.

Figure 8. F_1-Scores per dataset and size of testset

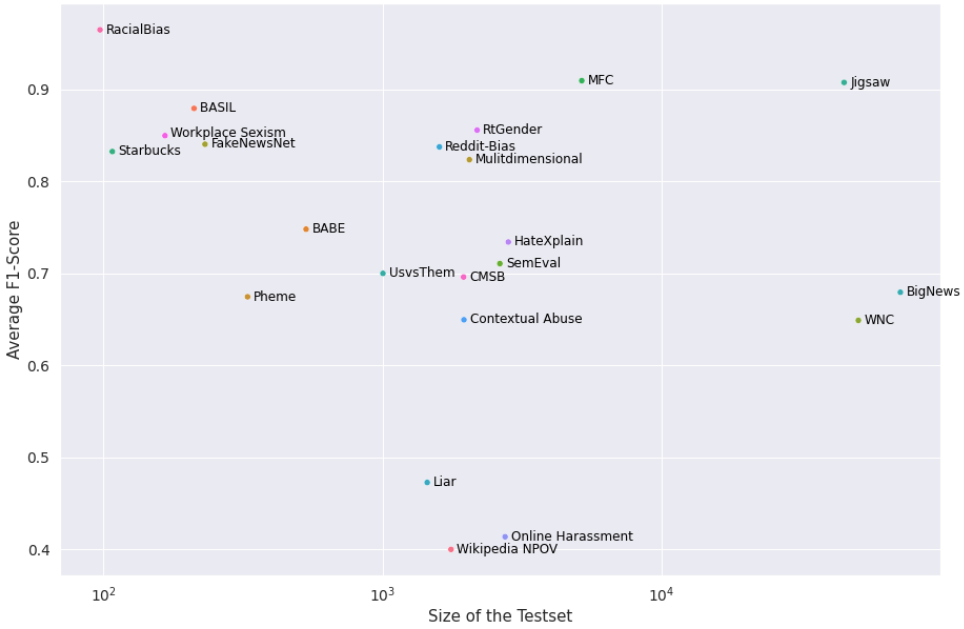


Figure 8 displays the size of the individual datasets' test sets and the performance (excluding the Wasseem Dataset). The test set size and performance are averaged over the five folds. If a dataset appeared in multiple tasks, the values are also averaged. Even though only the test set size is displayed, the size of the test set directly correlates with the size of the training set. A positive linear relationship between dataset size and performance is not visible. On the contrary, the small datasets seem to have high performance. This performance might confirm the exploratory impression that a lot of the smaller datasets, which are manually annotated, are often of better quality and contain, therefore, less noise than some of the larger datasets. If such a trade-o exists, that smaller datasets contain less noise than bigger ones, the concerns raised about the di ering dataset size play only a subordinate role. However, it is noteworthy that this does not mean such an e ect does not exist in the benchmark.

6.5 A weighted score

The overall scores displayed in Table 6 are calculated as direct F_1-Scores on all predictions on the test set. This approach is similar to a micro-average score, where

a global score is calculated out of the sum of true/false positives/negatives of all classes. However, it is not the typical micro-average, as the actual classes consist of whether a statement is biased or unbiased. Nevertheless, since the impact of the associated datasets is of interest, this score will be called the micro-average score. Alternatively to the micro-average score, a macro-average score is often calculated [Manning et al., 2008]. Instead of summing up results from all classes, an individual score is calculated for each class. The final overall score is then an average of these individual scores. Transferred to the situation at hand, this would imply first calculating an F_1-Score for every dataset individually and then averaging them.[1] Since it neglects the size of datasets, the macro-average score is potentially more balanced. Table 8 displays the macro-average F_1-Scores with regard to the datasets.

Table 8. Macro-average F_1-Scores by datasets

	Linguistic Bias			Cognitive Bias			Text Level Context Bias			Hate Speech	
	Model	F_1-Score		Model	F_1-Score		Model	F_1-Score		Model	F_1-Score
1.	Bart	0.7664	1.	ConvBERT	0.4995	1.	ConvBERT	0.7532	1.	Bart	0.7310
2.	RoBERTa-Twitter	0.7479	2.	RoBERTa-Twitter	0.4986	2.	Bart	0.7477	2.	ConvBERT	0.7248
3.	GPT-2	0.7459	3.	GPT-2	0.4968	3.	RoBERTa-Twitter	0.7382	3.	RoBERTa-Twitter	0.7229
4.	ConvBERT	0.7283	4.	ELECTRA	0.4949	4.	ELECTRA	0.7347	4.	GPT-2	0.7198
5.	ELECTRA	0.7136	5.	Bart	0.4881	5.	GPT-2	0.7075	5.	ELECTRA	0.7184

	Gender Bias			Racial Bias			Fake News			Political Bias	
	Model	F_1-Score		Model	F_1-Score		Model	F_1-Score		Model	F_1-Score
1.	ELECTRA	0.8211	1.	ELECTRA	0.6170	1.	RoBERTa-Twitter	0.7533	1.	ConvBERT	0.7110
2.	Bart	0.8168	2.	ConvBERT	0.6153	2.	ConvBERT	0.7382	2.	RoBERTa-Twitter	0.7036
3.	ConvBERT	0.8119	3.	Bart	0.6103	3.	Bart	0.7236	3.	ELECTRA	0.6989
4.	RoBERTa-Twitter	0.8116	4.	RoBERTa-Twitter	0.6070	4.	ELECTRA	0.7049	4.	GPT-2	0.6909
5.	GPT-2	0.7852	5.	GPT-2	0.5961	5.	GPT-2	0.6596	5.	Bart	0.6804

The biggest difference between micro and macro-average is at the cognitive and racial bias tasks. For both tasks, the micro (on average cognitive: 0.70 racial: 0.87) is much higher than the macro-average score (on average cognitive: 0.49 racial: 0.61). After looking at Figure 7 and Figure 8 this is no surprise. The BigNews Corpus showed a much higher performance for cognitive bias than the Liar Dataset. The BigNews Corpus being substantially bigger most likely skewed the overall score in favor of the BigNews Corpus's performance. Racial bias is mainly explained by the problems with Waseem and Hovy [2016]'s dataset. Excluding it results in the score remaining almost unchanged. For Hate Speech detection, the macro-average score also lies well below the micro score (macro on average: 0.88, micro: 0.72). One reason for this difference is the big Jigsaw dataset losing influence, which received a high score. So, also here, one big dataset had a decisive influence on the overall score. For linguistic, text-level context, gender, and political bias, the scores remain relatively stable. A higher score can be seen only for the remaining task, fake news

[1] Another alternative would be a weighted score. For this score, individual F_1-Scores are calculated for each class (as for the macro-average). Instead of a simple average, these scores are multiplied by the class share of the whole dataset and then summed up. So if a dataset consists of 60% of one class, this class's score makes up 60% of the final score. The weighted score is not useful for identifying whether individual datasets greatly influence the overall score.

detection (on average micro: 0.66, macro: 0.71). Here the biggest and worst performing Liar Dataset explains the difference.

In summary, for half of the scores, the method of calculating the score makes a substantial difference. The micro score has the advantage that it is easier to implement and understand. Mainly because the datasets are different from the actual classes and the transfer to the datasets is unconventional. Furthermore, it better reflects the actual performance of the dataset. However, assuming each dataset covers an equally important aspect of the bias task it belongs to, the macro score offers a much more balanced result. Seeing that the scores are mainly decreased, it also seems to be the more conservative score. Both scores can be useful. So, stating both will increase the informative value of MBIB's results.

CHAPTER 7
Limitations and Outlook

7.1 Theoretical Restrictions

In order to have a comprehensive benchmark covering all of media bias, the tasks of MBIB must cover as many aspects of media bias as possible. Aside from those tasks created, other bias tasks could be considered in the benchmark. Such tasks include, for example, religious bias, describing the discrimination or misrepresentation based on an individual's religion [Manzini et al., 2019]. Basing the task selection on societal relevance and high research interest demands a continuous consideration of which tasks to include. Taking the framework of Spinde et al. [2022a] as the basis for the media bias-inducing tasks offers the advantage of having tasks that conceptualize media bias in an all-encompassing way. It does, however, require the tasks to be more intuitive to understand. It often remains unclear where an occurring bias fits in as, e.g., certain lexical features are usually concomitants of phrasing bias (a bias belonging to Text-Level Context Bias).

7.2 Data Limitations

A lack of high-quality data still limits the media bias benchmark. This shortage is especially problematic in the media bias-inducing tasks. There are insufficient datasets for reporting-level bias to be included as a task. However, also for cognitive bias and text-level context bias, the data foundation is with only two datasets

less diverse than in the other tasks. Combined worse performance of the datasets used in cognitive bias shows that they might not fit well together. Only the creation of task-specific datasets can overcome this limitation. Also, critically reevaluating what alternative datasets might be used in these tasks might improve them.

Another critical consideration is needed on the heavy dependency on social media data. Since some tasks need more corpora based on news articles, using them could not be avoided. However, it becomes a problem for MBIB if it wants to claim that a model that, e.g., performs well on the racial bias task is a well-suited model to detect racial bias as part of media coverage. This limitation is important as it can be expected that the bias present in social media differs from those found in the media. In the media, the bias is often more subtle. Racial bias on social media is, for example, often accompanied by slurs and defamation. Something unlikely to be found so explicitly in news articles.

The largest dataset is based on distant labeling. For BigNews [Liu et al., 2022], labels for articles are defined by the corresponding allsides.com outlet classification. Such distant labeling techniques are likely to include much noise. Ganguly et al. [2020] find that the political leaning of the news outlet does not necessarily translate into the leaning of an individual article of that outlet. However, the models' performance on the distantly labeled datasets is not considerably lower than on manually labeled datasets. This performance might be influenced by the distantly labeled datasets being substantially bigger.

Many more limitations concerning individual dataset qualities can be named. Especially because there are big differences in factors like how many annotators are used or how well the annotators are trained. Further differences exist in how high the inter-annotator agreement is and how missing values and non-agreements are handled. Different creators of datasets having different definitions of what constitutes, e.g., hate speech, makes combining the datasets even more challenging.

The dataset overview and benchmark can also serve as a survey on the current status of data availability in media bias research. It shows that dataset creation focuses on a few aspects of media bias. There is an abundance of datasets on the political leaning of articles or bias through linguistic features, but none on, e.g., racial bias in news articles.

7.3 Experimental considerations

When setting model baselines for each of MBIB's tasks, there are limitations on how many models could be trained and tested. The proxy task represented an attempt to place the selection on an empirical foundation. The surprisingly high

performance of GPT-2 on the proxy task but relatively low performance in the experiments indicates that the generalizability of the proxy task might be limited. The models used do, however, cover a wide range of models. They included autoencoding (RoBERTa-Twitter, ConvBERT, ELECTRA), autoregressive (GPT-2), and sequence-to-sequence (BART) models, and a specialized pretraining paradigm (RoBERTa-Twitter). Due to the high amount of social media training data, this model might have an overrepresented advantage. Even though the proxy task might generalize insufficiently, the general finding that the model choice seems to have little influence on the performance makes it doubtful whether training and testing more transformer models with the same experimental setup would change the findings.

There is, however, room for improving the models' performance within the model training. So far, only the number of epochs is optimized, while other hyperparameters like the learning rate, batch size, and dropout rate are set to best practice values from the literature. Hyperparameter tuning does offer the potential for improvement here. Another important critical aspect of the experimental, but also MBIB's, design is the combination of multiple datasets into one task. Even though the experiment showed that larger datasets do not perform better than smaller ones, the sizes of datasets remain highly imbalanced. It remains to be seen what effect this imbalance has on the performance scores.

7.4 Future work

In order to make the media bias dataset collection, and the MBIB resources available to other researchers, a platform is needed where access can be granted as easily as possible. Such a platform (for example, Huggingface Spaces) would also allow associating visualizations aimed at understanding the structure and properties of the datasets and not included in this thesis. Ideally, on such a platform, the data for each task is available in the unified format with a non-public testing set as implemented by Wang et al. [2019b] for GLUE. Standing in the way of publishing the data is the problem that the data is not generally public for two datasets, but non-disclosure agreements had to be signed to get them from their creators. For Twitter datasets, the actual tweets may not be published but only tweet IDs. An agreement with the creators to publish them is needed, or instructions for access to the data and preprocessing scripts could be provided. Giving only instructions would, however, complicate MBIB's usage.

The datasets of each task leave much information not used in the model baselines. For instance, multi-categorical or numerical data is collapsed into a binary format. Additionally, e.g., BABE [Spinde et al., 2021b] does not only offer information on whether a sentence is biased but also which words are inducing the bias. It could be tested whether using such additional information would increase the

models' detection ability. Additional analyses on the similarity of datasets and the effects of training models on multiple datasets might yield further insights into the properties of MBIB. Even though duplicates are removed, finding and applying a similarity metric could reveal whether there are highly related datasets or structural differences within or between datasets. For the latter, a model could be trained on all task datasets separately and on the combined data. The performance difference can then indicate positive or negative transfers that the combination of datasets has.

Testing transformer models on particular tasks only sets the baseline for model performances in media bias. Other models and training objectives could lead to findings on improving the detection. Even though, as shown in previous sections, the research currently focuses on transformer models, other models like CNNs or LSTMs are usable for media bias detection. Combining data from multiple tasks in multi-task approaches will likely improve the performance on individual tasks [Aribandi et al., 2022]. MBIB would further enable research on the capabilities of few-shot learning in media bias detection. It could help to see whether models trained for a specific bias can detect other kinds of bias after being fine-tuned on only a few examples of this bias [Wang et al., 2020]. Especially since high-quality datasets are rare for some tasks, few-shot learning has big potential in media bias research.

Due to this limitation of datasets and as shown in section 7.2, future research should also concentrate on the creation of more high-quality datasets. The creation should focus on tasks with limited current availability. These tasks are primarily those derived from the framework by Spinde et al. [2022a] which tries to comprehensively cover media bias. A departure away from singular independent research fields towards wider tasks would enable also wider detection. When focussing on a singular field the models might only be able to answer, e.g., "Is there Gender Bias in this article?" while wider tasks would allow models to answer "Is there bias in this article?" (and ideally be able to identify why and where there is a bias).

CHAPTER 8

Conclusion

At the beginning of this thesis stands the question of what model would be the best to use when detecting media bias. However, a benchmark and a comparison framework are needed to answer this question. Something that did not yet exist in media bias research. The thesis, therefore, sets out to create such a benchmark and framework.

To create the benchmark, MBIB, media bias tasks are chosen and justified, and datasets for the task are found. The dataset collection constructed for this purpose can help other researchers to help find suitable datasets. However, it also points out gaps in the research and calls for dataset creation, especially for tasks like reporting-level bias and racial bias.

Having created the benchmark tasks and collected pertinent data, an evaluation framework enables the comparison of models on the benchmark. After selecting five transformer models via a proxy task, their performances on the eight MBIB tasks are compared. The results show that there are big differences in performance between tasks. While there are also differences between models within one task, no clear best model emerged. GPT-2 and ELECTRA seem to be the least suitable, and RoBERTa-Twitter and ConvBERT are best suitable regardless of the micro or macro-average F_1-Score. Generally, the transformer model choice matters less than initially assumed. Special attention is paid to how the combination of different datasets influences the result. While dataset size does not impact the score, with the micro and macro scores, two different metrics are introduced to account for this. The results of the two scores are similar.

The results can be taken as a basis for an informed model choice when choosing a model for media bias detection and therefore fills the gap identified at the beginning. However, the search for the best suitable model for media bias detection is still ongoing. There are many other paradigms like multi-task, meta, or few-shot learning that could be tested on MBIB. Therefore, future research is encouraged to test their models on MBIB and raise the benchmark.

Having a well-performing model that reliably can detect media bias can have a significant impact on media consumption and creation. On the consumption side, it could help readers identify instances of misinformation or other biased coverage they read. It could help educate readers and lead to a better-informed audience. On the creation side, it could help authors to improve the accuracy and objectivity

of their writing. It is, however, important to also keep in mind possible negative consequences. Models that control media production or what readers consume have the potential to interfere with fundamental rights of freedom of the press and freedom of speech. Furthermore, constant dialogue will be necessary to determine what counts as biased. Definitions differ not only over time but also geographically and between cultures. Taking such differences into account could be a potential next improvement in media bias research.

The rise of improved transformer-based sequence-to-sequence models such as ChatGPT [Radford et al., 2019] offers further potential for media bias detection. Models like this could detect bias and reason why a statement is biased and reformulate it. Giving reasonings behind a decision will make it much more likely that audiences will accept the classification of a model (as opposed to, e.g., only seeing the label "biased" next to an article). However, also these models will need to be checked on whether their classifications are correct. To do this, they can easily be tested on the MBIB framework.

While there are many future possibilities for improving media bias detection, they all require a systematic benchmark which this thesis hopes to offer. To improve MBIB, further research and resource creation on previously neglected fields are necessary. All combined with a critical discourse on how to evaluate technology's ever quicker emerging possibilities.

Bibliography

Oluseyi Adegbola, Jacqueline Skarda-Mitchell, and Sherice Gearhart. Everything's negative about Nigeria: A study of US media reporting on Nigeria. 2018. URL https://journals.sagepub.com/doi/full/10.1177/1742766518760086.

Jigsaw/Conversation AI. Jigsaw unintended bias in toxicity classification., March 2019. URL https://www.kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification/data.

Davide Anguita, Luca Ghelardoni, Alessandro Ghio, Luca Oneto, and Sandro Ridella. The 'K'in K-fold cross validation. In *20th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, pages 441–446. 2012.

Alberto Ardèvol-Abreu and Homero Gil de Zúñiga. E□ects of Editorial Media Bias Perception and Media Trust on the Use of Traditional, Citizen, and Social Media News. *Journalism & Mass Communication Quarterly*, 94(3):703–724, September 2017. ISSN 1077-6990. doi: 10.1177/1077699016654684. URL https://doi.org/10.1177/1077699016654684. Publisher: SAGE Publications Inc.

Vamsi Aribandi, Yi Tay, Tal Schuster, Jinfeng Rao, Huaixiu Steven Zheng, Sanket Vaibhav Mehta, Honglei Zhuang, Vinh Q. Tran, Dara Bahri, Jianmo Ni, Jai Gupta, Kai Hui, Sebastian Ruder, and Donald Metzler. ExT5: Towards Extreme Multi-Task Scaling for Transfer Learning, January 2022. URL http://arxiv.org/abs/2111.10952. arXiv:2111.10952 [cs].

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.148. URL https://www.aclweb.org/anthology/2020.findings-emnlp.148.

Soumya Barikeri, Anne Lauscher, Ivan Vulić, and Goran Glavaš. 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, 2021.

Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.151. URL https://aclanthology.org/2021.acl-long.151.

Eric Baumer, Elisha Elovic, Ying Qin, Francesca Polletta, and Geri Gay. 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, 2015. Association for Computational Linguistics. doi: 10.3115/v1/N15-1171. URL http://aclweb.org/anthology/N15-1171.

Lucy Bennett, Susan Bison, Marina Morani, Lorena Riveiro Rodríguez, Laura Pomarius, Sandra Kaulfuss, and Isabel Sundberg. Report authors: Mike Berry, Inaki Garcia-Blanco, Kerry Moore.

Daniel Berrar. Cross-validation, 2019.

Camiel J. Beukeboom and Christian Burgers. Linguistic Bias. In *Oxford Research Encyclopedia of Communication*. Oxford University Press, July 2017. ISBN 978-0-19-022861-3. doi: 10.1093/acrefore/9780190228613.013.439. URL https://oxfordre.com/communication/view/10.1093/acrefore/9780190228613.001.0001/acrefore-9780190228613-e-439.

Christopher M. Bishop. *Neural Networks for Pattern Recognition*. Oxford university press, 1995. URL https://books.google.de/books?hl=de&lr=&id=T0S0BgAAQBAJ&oi=fnd&pg=PP1&dq=neural+Networks+for+Pattern+Recognition&ots=jN9TsE7vsh&sig=kktyGirTWCe7xJOZ7NmVI18j6AQ#v=onepage&q=neural%20Networks%20for%20Pattern%20Recognition&f=false.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, and Tom Henighan. Language Models are Few-Shot Learners. *NeurIPS*, page 25, 2020.

Ceren Budak, Sharad Goel, and Justin M. Rao. Fair and Balanced? Quantifying Media Bias through Crowdsourced Content Analysis. *Public Opinion Quarterly*, 80 (S1):250–271, 2016. ISSN 0033-362X, 1537-5331. doi: 10.1093/poq/nfw007. URL https://academic.oup.com/poq/article-lookup/doi/10.1093/poq/nfw007.

Camly Bui. How Online Gatekeepter Guard our View - News Portal's Inclusion and Ranking of Media News and Events. *Global Media Journal*, 9(16), 2010.

Wei-Fan Chen, Khalid Al Khatib, Benno Stein, and Henning Wachsmuth. Detecting Media Bias in News Articles using Gaussian Bias Distributions. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4290–4300, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.383. URL https://www.aclweb.org/anthology/2020.findings-emnlp.383.

Lamogha Chiazor, Geeth R. De Mel, Graham White, Gwilym Newton, Joe Pavitt, and Richard Tomsett. An Automated Framework to Identify and Eliminate Systemic Racial Bias in the Media. February 2021.

Nancy Chinchor. MUC-4 evaluation metrics. In *Proceedings of the 4th conference on Message understanding*, MUC4 '92, pages 22–29, USA, June 1992. Association for Computational Linguistics. ISBN 978-1-55860-273-1. doi: 10.3115/1072064.1072067. URL https://doi.org/10.3115/1072064.1072067.

Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. The echo chamber e□ect on social media. *Proceedings of the National Academy of Sciences*, 118(9):e2023301118, March 2021. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2023301118. URL https://pnas.org/doi/full/10.1073/pnas.2023301118.

Kevin Clark, Minh-Thang Luong, Quoc Le, and Christopher Manning. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. 2020. URL https://openreview.net/forum?id=r1xMH1BtvB.

Alexis Conneau and Douwe Kiela. SentEval: An Evaluation Toolkit for Universal Sentence Representations, March 2018. URL http://arxiv.org/abs/1803.05449. arXiv:1803.05449 [cs].

Marta R. Costa-jussà. An analysis of gender bias studies in natural language processing. *Nature Machine Intelligence*, 1(11):495–496, November 2019. ISSN 2522-5839. doi: 10.1038/s42256-019-0105-5. URL https://www.nature.com/articles/s42256-019-0105-5. Number: 11 Publisher: Nature Publishing Group.

Jamell Dacon and Haochen Liu. 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, June 2021. Association for Computing Machinery. ISBN 978-1-4503-8313-4. doi: 10.1145/3442442.3452325. URL https://doi.org/10.1145/3442442.3452325.

Dave D'Alessio and Mike Allen. Media Bias in Presidential Elections: A Meta-Analysis. *Journal of Communication*, 50(4):133–156, December 2000. ISSN 0021-9916, 1460-2466. doi: 10.1111/j.1460-2466.2000.tb02866.x. URL https://academic.oup.com/joc/article/50/4/133-156/4110147.

Stefano DellaVigna and Ethan Kaplan. The Fox News E□ect: Media Bias and Voting*. *The Quarterly Journal of Economics*, 122(3):1187–1234, August 2007. ISSN 0033-5533. doi: 10.1162/qjec.122.3.1187. URL https://doi.org/10.1162/qjec.122.3.1187.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North*, pages 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL http://aclweb.org/anthology/N19-1423.

N. A. Diamantidis, D. Karlis, and E. A. Giakoumakis. Unsupervised stratification of cross-validation for accuracy estimation. *Artificial Intelligence*, 116(1):1–16, January 2000. ISSN 0004-3702. doi: 10.1016/S0004-3702(99)00094-6. URL https://www.sciencedirect.com/science/article/pii/S0004370299000946.

Marko Dragojevic, Alexander Sink, and Dana Mastro. Evidence of Linguistic Intergroup Bias in U.S. Print News Coverage of Immigration. *Journal of Language and Social Psychology*, 36(4):462–472, September 2017. ISSN 0261-927X, 1552-6526. doi: 10.1177/0261927X16666884. URL http://journals.sagepub.com/doi/10.1177/0261927X16666884.

Kristin Nicole Dukes and Sarah E. Gaither. Black Racial Stereotypes and Victim Blaming: Implications for Media Coverage and Criminal Proceedings in Cases of Police Violence against Racial and Ethnic Minorities. *Journal of Social Issues*, 73(4):789–807, 2017. ISSN 1540-4560. doi: 10.1111/josi.12248. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/josi.12248. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/josi.12248.

Robert M. Entman. Framing Bias: Media in the Distribution of Power. *Journal of Communication*, 57(1):163–173, March 2007. ISSN 00219916, 14602466. doi: 10.1111/j.1460-2466.2006.00336.x. URL https://academic.oup.com/joc/article/57/1/163-173/4102665.

Jo Ellen Fair. War, Famine, and Poverty: Race in the Construction of Africa's Media Image. *Journal of Communication Inquiry*, 17(2):5–22, July 1993. ISSN 0196-8599. doi: 10.1177/019685999301700202. URL https://doi.org/10.1177/019685999301700202. Publisher: SAGE Publications Inc.

Lisa Fan, Marshall White, Eva Sharma, Ruisi Su, Prafulla Kumar Choubey, Ruihong Huang, and Lu Wang. In Plain Sight: Media Bias Through the Lens of Factual Reporting, September 2019. URL http://arxiv.org/abs/1909.02670. arXiv:1909.02670 [cs].

Ibrahim Farha and Walid Magdy. Benchmarking transformer-based language models for Arabic sentiment and sarcasm detection. In *Proceedings of the sixth Arabic natural language processing workshop*, pages 21–31, 2021.

Stanley Feldman. Political ideology. In *The Oxford handbook of political psychology, 2nd ed*, pages 591–626. Oxford University Press, New York, NY, US, 2013. ISBN 978-0-19-976010-7.

Paula Fortuna and Sérgio Nunes. A Survey on Automatic Detection of Hate Speech in Text. *ACM Computing Surveys*, 51(4):85:1–85:30, July 2018. ISSN 0360-0300. doi: 10.1145/3232676. URL https://doi.org/10.1145/3232676.

Tadayoshi Fushiki. Estimation of prediction error by using K-fold cross-validation. *Statistics and Computing*, 21(2):137–146, April 2011. ISSN 1573-1375. doi: 10.1007/s11222-009-9153-8. URL https://doi.org/10.1007/s11222-009-9153-8.

Michael Färber, Victoria Burkard, Adam Jatowt, and Sora Lim. A Multidimensional Dataset Based on Crowdsourcing for Analyzing and Detecting News Bias. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 3007–3014, Virtual Event Ireland, October 2020. ACM. ISBN 978-1-4503-6859-9. doi: 10.1145/3340531.3412876. URL https://dl.acm.org/doi/10.1145/3340531.3412876.

Soumen Ganguly, Juhi Kulshrestha, Jisun An, and Haewoon Kwak. Empirical Evaluation of Three Common Assumptions in Building Political Media Bias Datasets. *Proceedings of the International AAAI Conference on Web and Social Media*, 14: 939–943, May 2020. ISSN 2334-0770, 2162-3449. doi: 10.1609/icwsm.v14i1.7362. URL https://ojs.aaai.org/index.php/ICWSM/article/view/7362.

Bertram Gawronski. Partisan bias in the identification of fake news. *Trends in Cognitive Sciences*, 25(9):723–724, September 2021. ISSN 13646613. doi: 10.1016/j.tics.2021.05.001. URL https://linkinghub.elsevier.com/retrieve/pii/S1364661321001224.

Matthew Gentzkow and Jesse M. Shapiro. Media Bias and Reputation. *Journal of Political Economy*, 114(2):280–316, April 2006. ISSN 0022-3808, 1537-534X. doi: 10.1086/499414. URL https://www.journals.uchicago.edu/doi/10.1086/499414.

Alan S Gerber, Dean Karlan, and Daniel Bergan. Does the Media Matter? A Field Experiment Measuring the E□ect of Newspapers on Voting Behavior and Political Opinions. *American Economic Journal: Applied Economics*, 1(2):35–52, March 2009. ISSN 1945-7782, 1945-7790. doi: 10.1257/app.1.2.35. URL https://pubs.aeaweb.org/doi/10.1257/app.1.2.35.

Torumoy Ghoshal. Racial Bias Twitter, 2018. URL https://github.com/tgh499/racial_bias_twitter.

Trystan S. Goetze and Darren Abramson. Bigger Isn't Better: The Ethical and Scientific Vices of Extra-Large Datasets in Language Models. In *13th ACM Web*

Science Conference 2021, pages 69–75, Virtual Event United Kingdom, June 2021. ACM. ISBN 978-1-4503-8525-1. doi: 10.1145/3462741.3466809. URL https://dl.acm.org/doi/10.1145/3462741.3466809.

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. Ho□man, 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. 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, June 2017. ACM. ISBN 978-1-4503-4896-6. doi: 10.1145/3091478.3091509. URL https://dl.acm.org/doi/10.1145/3091478.3091509.

Hila Gonen and Yoav Goldberg. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. 2019. doi: 10.48550/ARXIV.1903.03862. URL https://arxiv.org/abs/1903.03862. Publisher: arXiv Version Number: 2.

Google. Perspective API, October 2022. URL https://www.perspectiveapi.com/.

Stephan Greene and Philip Resnik. More than Words: Syntactic Packaging and Implicit Sentiment. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 503–511, Boulder, Colorado, June 2009. Association for Computational Linguistics. URL https://aclanthology.org/N09-1057.

Dylan Grosz and Patricia Conde-Cespedes. 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, 2020. ISBN 978-3-030-60469-1 978-3-030-60470-7. doi: 10.1007/978-3-030-60470-7_11. URL https://link.springer.com/10.1007/978-3-030-60470-7_11.

Shijia Guo and Kenny Q. Zhu. Modeling Multi-level Context for Informational Bias Detection by Contrastive Learning and Sentential Graph Network, January 2022. URL http://arxiv.org/abs/2201.10376. arXiv:2201.10376 [cs].

Yuting Guo, Xiangjue Dong, Mohammed Ali Al-Garadi, Abeed Sarker, Cecile Paris, and Diego Molla Aliod. Benchmarking of transformer-based pre-trained models on social media text classification datasets. In *Proceedings of the The 18th Annual Workshop of the Australasian Language Technology Association*, pages 86–91, 2020.

Felix Hamborg, Karsten Donnay, and Bela Gipp. Automated identification of media bias in news articles: an interdisciplinary literature review. *International Journal on Digital Libraries*, 20(4):391–415, December 2019. ISSN 1432-5012, 1432-1300. doi: 10.1007/s00799-018-0261-y. URL http://link.springer.com/10.1007/s00799-018-0261-y.

Roderick P. Hart, Kathleen J. Turner, and Ralph E. Knupp. Religion and the Rhetoric of the Mass Media. *Review of Religious Research*, 21(3):256, 1980. ISSN 0034673X. doi: 10.2307/3509807. URL https://www.jstor.org/stable/3509807?origin=crossref.

Gerald Ki Wei Huang and Jun Choi Lee. Hyperpartisan News and Articles Detection Using BERT and ELMo. In *2019 International Conference on Computer and Drone Applications (IConDA)*, pages 29–32, Kuching, Malaysia, December 2019. IEEE. ISBN 978-1-72816-592-9 978-1-72816-593-6. doi: 10.1109/IConDA47345.2019.9034917. URL https://ieeexplore.ieee.org/document/9034917/.

Christoph Hube and Besnik Fetahu. Neural Based Statement Classification for Biased Language. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 195–203, Melbourne VIC Australia, January 2019. ACM. ISBN 978-1-4503-5940-5. doi: 10.1145/3289600.3291018. URL https://dl.acm.org/doi/10.1145/3289600.3291018.

Pere-Lluís Huguet Cabot, David Abadi, Agneta Fischer, and Ekaterina Shutova. 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, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.165. URL https://aclanthology.org/2021.eacl-main.165.

Laura Jacobs, Alyt Damstra, Mark Boukes, and Knut De Swert. Back to Reality: The Complex Relationship Between Patterns in Immigration News Coverage and Real-World Developments in Dutch and Flemish Newspapers (1999–2015). *Mass Communication and Society*, 21(4):473–497, July 2018. ISSN 1520-5436, 1532-7825. doi: 10.1080/15205436.2018.1442479. URL https://www.tandfonline.com/doi/full/10.1080/15205436.2018.1442479.

Zihang Jiang, Weihao Yu, Daquan Zhou, Yunpeng Chen, Jiashi Feng, and Shuicheng Yan. ConvBERT: Improving BERT with Span-based Dynamic Convolution. 2020. doi: 10.48550/ARXIV.2008.02496. URL https://arxiv.org/abs/2008.02496. Publisher: arXiv Version Number: 3.

Katikapalli Subramanyam Kalyan, Ajit Rajasekharan, and Sivanesan Sangeetha. AMMUS : A Survey of Transformer-based Pretrained Models in Natural Language Processing. 2021. doi: 10.48550/ARXIV.2108.05542. URL https://arxiv.org/abs/2108.05542. Publisher: arXiv Version Number: 2.

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

Brent Kitchens, Steve L. Johnson, and Peter Gray. Understanding Echo Chambers and Filter Bubbles: The Impact of Social Media on Diversification and Partisan Shifts in News Consumption. *MIS Quarterly*, 44(4):1619–1649, December 2020. ISSN 02767783, 21629730. doi: 10.25300/MISQ/2020/16371.

Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. All-in-one: Multi-task Learning for Rumour Verification. 2018. doi: 10.48550/ARXIV.1806.03713. URL https://arxiv.org/abs/1806.03713.

Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of 14th International Joint Conference on AI*, pages 1137–45, 1995.

Anne C. Kroon, Damian Trilling, and Tamara Raats. Guilty by Association: Using Word Embeddings to Measure Ethnic Stereotypes in News Coverage. *Journalism & Mass Communication Quarterly*, 98(2):451–477, June 2021. ISSN 1077-6990. doi: 10.1177/1077699020932304. URL https://doi.org/10.1177/1077699020932304. Publisher: SAGE Publications Inc.

Haewoon Kwak, Jisun An, and Yong-Yeol Ahn. A Systematic Media Frame Analysis of 1.5 Million New York Times Articles from 2000 to 2017. In *12th ACM Conference on Web Science*, pages 305–314, Southampton United Kingdom, July 2020. ACM. ISBN 978-1-4503-7989-2. doi: 10.1145/3394231.3397921. URL https://dl.acm.org/doi/10.1145/3394231.3397921.

Lesley Lavery. 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):877–910, December 2013. ISSN 15555623. doi: 10.1111/polp.12051. URL https://onlinelibrary.wiley.com/doi/10.1111/polp.12051.

Konstantia Lazaridou and Ralf Krestel. Identifying Political Bias in News Articles. *Bull. IEEE TEch. Comm. Digit. Libr.*, (12), 2016.

Nayeon Lee, Yejin Bang, Andrea Madotto, and Pascale Fung. Mitigating Media Bias through Neutral Article Generation. 2021. doi: 10.48550/ARXIV.2104.00336. URL https://arxiv.org/abs/2104.00336. Publisher: arXiv Version Number: 1.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denois-

ing Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.703. URL https://www.aclweb.org/anthology/2020.acl-main.703.

Yanling Li, Guoshe Sun, and Yehang Zhu. Data Imbalance Problem in Text Classification. In *2010 Third International Symposium on Information Processing*, pages 301–305, October 2010. doi: 10.1109/ISIP.2010.47.

Sora Lim, Adam Jatowt, and Y Masatoshi. Creating a dataset for fine-grained bias detection in news articles. 12, pages 1–35, 2020.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. 2019. doi: 10.48550/ARXIV. 1907.11692. URL https://arxiv.org/abs/1907.11692. Publisher: arXiv Version Number: 1.

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

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval*. Cambridge university press, Cambridge, 2008. ISBN 978-0-521-86571-5.

Thomas Manzini, Yao Chong Lim, Yulia Tsvetkov, and Alan W. Black. Black is to Criminal as Caucasian is to Police: Detecting and Removing Multiclass Bias in Word Embeddings, July 2019. URL http://arxiv.org/abs/1904.04047. arXiv:1904.04047 [cs, stat].

Binny Mathew, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. Spread of Hate Speech in Online Social Media. In *Proceedings of the 10th ACM Conference on Web Science*, pages 173–182, Boston Massachusetts USA, June 2019. ACM. ISBN 978-1-4503-6202-3. doi: 10.1145/3292522.3326034. URL https://dl.acm.org/doi/10.1145/3292522.3326034.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14867–14875, May 2021. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v35i17.17745. URL https://ojs.aaai.org/index.php/AAAI/article/view/17745.

Leeja Mathew and V R Bindu. A Review of Natural Language Processing Techniques for Sentiment Analysis using Pre-trained Models. In *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, pages 340–345, Erode, India, March 2020. IEEE. ISBN 978-1-72814-889-2. doi: 10.1109/ICCMC48092.2020.ICCMC-00064. URL https://ieeexplore.ieee.org/document/9076502/.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. Mixed Precision Training, February 2018. URL http://arxiv.org/abs/1710.03740. arXiv:1710.03740 [cs, stat].

Tomas Mikolov, Kai Chen, Greg Corrado, and Je□rey Dean. E□cient Estimation of Word Representations in Vector Space. 2013. doi: 10.48550/ARXIV.1301.3781. URL https://arxiv.org/abs/1301.3781. Publisher: arXiv Version Number: 3.

Seong-Jae Min and John C. Feaster. Missing Children in National News Coverage: Racial and Gender Representations of Missing Children Cases. *Communication Research Reports*, 27(3):207–216, August 2010. ISSN 0882-4096. doi: 10.1080/08824091003776289. URL https://doi.org/10.1080/08824091003776289. Publisher: Routledge _eprint: https://doi.org/10.1080/08824091003776289.

Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. Hate speech detection and racial bias mitigation in social media based on BERT model. *PLOS ONE*, 15(8):e0237861, August 2020. ISSN 1932-6203. doi: 10.1371/journal.pone.0237861. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0237861. Publisher: Public Library of Science.

Sendhil Mullainathan and Andrei Shleifer. Media Bias. Technical Report w9295, National Bureau of Economic Research, Cambridge, MA, October 2002. URL http://www.nber.org/papers/w9295.pdf.

Raymond S. Nickerson. Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology*, 2(2):175–220, June 1998. ISSN 1089-2680, 1939-1552. doi: 10.1037/1089-2680.2.2.175. URL http://journals.sagepub.com/doi/10.1037/1089-2680.2.2.175.

Timoth Niven and Hung-Yu Kao. Measuring Alignment to Authoritarian State Media as Framing Bias - ACL Anthology. In *Proceedings of the 3rd NLP4IF Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 11–21, Barcelona, Spain, 2020. URL https://aclanthology.org/2020.nlp4if-1.2/.

Toussaint Nothias. How Western Journalists *Actually* Write About Africa: Reassessing the myth of representations of Africa. *Journalism Studies*, 19(8):1138–1159, June 2018. ISSN 1461-670X, 1469-9699. doi: 10.1080/1461670X.2016.

1262748. URL https://www.tandfonline.com/doi/full/10.1080/1461670X.2016.1262748.

Cristian Padurariu and Mihaela Elena Breaban. Dealing with Data Imbalance in Text Classification. *Procedia Computer Science*, 159:736–745, 2019. ISSN 18770509. doi: 10.1016/j.procs.2019.09.229. URL https://linkinghub.elsevier.com/retrieve/pii/S1877050919314152.

Thomas E. Patterson. News Coverage of the 2016 General Election: How the Press Failed the Voters, December 2016. URL https://papers.ssrn.com/abstract=2884837.

Je□rey Pennington, Richard Socher, and Christopher Manning. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL http://aclweb.org/anthology/D14-1162.

Alexandra Guedes Pinto, Henrique Lopes Cardoso, Isabel Margarida Duarte, Catarina Vaz Warrot, and Rui Sousa-Silva. Biased Language Detection in Court Decisions. In Cesar Analide, Paulo Novais, David Camacho, and Hujun Yin, editors, *Intelligent Data Engineering and Automated Learning – IDEAL 2020*, Lecture Notes in Computer Science, pages 402–410, Cham, 2020. Springer International Publishing. ISBN 978-3-030-62365-4. doi: 10.1007/978-3-030-62365-4_38.

Keval Pipalia, Rahul Bhadja, and Madhu Shukla. Comparative Analysis of Different Transformer Based Architectures Used in Sentiment Analysis. In *2020 9th International Conference System Modeling and Advancement in Research Trends (SMART)*, pages 411–415, Moradabad, India, December 2020. IEEE. ISBN 978-1-72818-908-6. doi: 10.1109/SMART50582.2020.9337081. URL https://ieeexplore.ieee.org/document/9337081/.

Lutz Prechelt. Early Stopping - But When? In Genevieve B. Orr and Klaus-Robert Müller, editors, *Neural Networks: Tricks of the Trade*, Lecture Notes in Computer Science, pages 55–69. Springer, Berlin, Heidelberg, 1998. ISBN 978-3-540-49430-0. doi: 10.1007/3-540-49430-8_3. URL https://doi.org/10.1007/3-540-49430-8_3.

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

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training., 2018.

Alec Radford, Je□rey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1 (8):9, 2019.

Shruti Raina. GENDER BIAS IN EDUCATION. *INTERNATIONAL JOURNAL OF RESEARCH PEDAGOGY AND TECHNOLOGY IN EDUCATION AND MOVEMENT SCIENCES*, 1(02), 2012. ISSN 2319-3050. URL https://ijems.net/index.php/ijem/article/view/10. Number: 02.

Hannah Rashkin, Sameer Singh, and Yejin Choi. Connotation Frames: A Data-Driven Investigation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 311–321, Berlin, Germany, 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1030. URL http://aclweb.org/anthology/P16-1030.

Marta Recasens, Christian Danescu-Niculescu-Mizil, and Dan Jurafsky. Linguistic models for analyzing and detecting biased language. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1650–1659, 2013.

Payam Refaeilzadeh, Lei Tang, and Huan Liu. Cross-Validation. In Ling Liu and M. Tamer Özsu, editors, *Encyclopedia of Database Systems*, pages 1–7. Springer, New York, NY, 2016. ISBN 978-1-4899-7993-3. doi: 10.1007/978-1-4899-7993-3_565-2. URL https://doi.org/10.1007/978-1-4899-7993-3_565-2.

Yasmim Mendes Rocha, Gabriel Acácio de Moura, Gabriel Alves Desidério, Carlos Henrique de Oliveira, Francisco Dantas Lourenço, and Larissa Deadame de Figueiredo Nicolete. 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*, October 2021. ISSN 1613-2238. doi: 10.1007/s10389-021-01658-z. URL https://doi.org/10.1007/s10389-021-01658-z.

Diego Saez-Trumper, Carlos Castillo, and Mounia Lalmas. Social media news communities: gatekeeping, coverage, and statement bias. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management - CIKM '13*, pages 1679–1684, San Francisco, California, USA, 2013. ACM Press. ISBN 978-1-4503-2263-8. doi: 10.1145/2505515.2505623. URL http://dl.acm.org/citation.cfm?doid=2505515.2505623.

Mattia Samory, Indira Sen, Julian Kohne, Fabian Floeck, and Claudia Wagner. "Call me sexist, but...": Revisiting Sexism Detection Using Psychological Scales and Adversarial Samples. 2020. doi: 10.48550/ARXIV.2004.12764. URL https://arxiv.org/abs/2004.12764.

Sihyeon Seong, Yegang Lee, Youngwook Kee, Dongyoon Han, and Junmo Kim. Towards Flatter Loss Surface via Nonmonotonic Learning Rate Scheduling. *UAI*, 2018.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 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, June 2020. ISSN 2167-6461, 2167-647X. doi: 10.1089/big.2020.0062. URL https://www.liebertpub.com/doi/10.1089/big.2020.0062.

Vivek K. Singh, Mary Chayko, Raj Inamdar, and Diana Floegel. 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):1281–1294, November 2020. ISSN 2330-1635, 2330-1643. doi: 10.1002/asi.24335. URL https://onlinelibrary.wiley.com/doi/10.1002/asi.24335.

Barea Sinno, Bernardo Oviedo, Katherine Atwell, Malihe Alikhani, and Junyi Jessy Li. Political Ideology and Polarization of Policy Positions: A Multidimensional Approach, May 2022. URL http://arxiv.org/abs/2106.14387. arXiv:2106.14387 [cs].

Deepak Soekhoe, Peter van der Putten, and Aske Plaat. On the Impact of Data Set Size in Transfer Learning Using Deep Neural Networks. In Henrik Boström, Arno Knobbe, Carlos Soares, and Panagiotis Papapetrou, editors, *Advances in Intelligent Data Analysis XV*, Lecture Notes in Computer Science, pages 50–60, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46349-0. doi: 10.1007/978-3-319-46349-0_5.

Wiktor Soral, Michał Bilewicz, and Mikołaj Winiewski. Exposure to hate speech increases prejudice through desensitization. *Aggressive Behavior*, 44(2):136–146, 2018. ISSN 1098-2337. doi: 10.1002/ab.21737. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/ab.21737. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ab.21737.

T. Spinde, L. Rudnitckaia, K. Sinha, F. Hamborg, B. Gipp, and K. Donnay. MBIC – A Media Bias Annotation Dataset Including Annotator Characteristics, May 2021a. URL http://arxiv.org/abs/2105.11910. arXiv:2105.11910 [cs].

Timo Spinde, Manuel Plank, Jan-David Krieger, Terry Ruas, Bela Gipp, and Akiko Aizawa. 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, November 2021b. doi: 10.18653/v1/2021.findings-emnlp.101. URL https://media-bias-research.org/wp-content/uploads/2022/01/Neural_Media_Bias_Detection_Using_Distant_Supervision_With_BABE___Bias_Annotations_By_Experts_MBG.pdf.

Timo Spinde, Lada Rudnitckaia, Jelena Mitrović, Felix Hamborg, Michael Granitzer, Bela Gipp, and Karsten Donnay. Automated identification of bias inducing words in news articles using linguistic and context-oriented

features. *Information Processing & Management*, 58(3):102505, January 2021c. ISSN 0306-4573. doi: https://doi.org/10.1016/j.ipm.2021.102505. URL https://www.sciencedirect.com/science/article/pii/S0306457321000157/pdfft?md5=64e81212b3bfa861d01a6fe3d5b979c3&pid=1-s2.0-S0306457321000157-main.pdf.

Timo Spinde, Smilla Hinterreiter, Fabian Haak, Helge Giese, Norman Meuschke, and Terry Ruas. Introducing the Media Bias Framework. An Interdisciplinary Literature Review on the Perception And Detection of Media Bias, 2022a. in review.

Timo Spinde, Jan-David Krieger, Terry Ruas, Jelena Mitrović, Franz Götz-Hahn, Akiko Aizawa, and Bela Gipp. Exploiting Transformer-based Multitask Learning for the Detection of Media Bias in News Articles. In *Proceedings of the iConference 2022*, Virtual event, March 2022b. doi: https://doi.org/10.1007/978-3-030-96957-8_20. URL https://media-bias-research.org/wp-content/uploads/2022/03/Spinde2022a_mbg.pdf.

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

Latanya Sweeney. Discrimination in Online Ad Delivery: Google ads, black names and white names, racial discrimination, and click advertising. *Queue*, 11(3):10–29, March 2013. ISSN 1542-7730. doi: 10.1145/2460276.2460278. URL https://doi.org/10.1145/2460276.2460278.

Edson C. Tandoc Jr. The facts of fake news: A research review. *Sociology Compass*, 13(9):e12724, 2019. ISSN 1751-9020. doi: 10.1111/soc4.12724. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/soc4.12724. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/soc4.12724.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. 2017. doi: 10.48550/ARXIV.1706.03762. URL https://arxiv.org/abs/1706.03762. Publisher: arXiv Version Number: 5.

Bertie Vidgen, Dong Nguyen, Helen Margetts, Patricia Rossini, and Rebekah Tromble. 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*, pages 2289–2303, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.182. URL https://aclanthology.org/2021.naacl-main.182.

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

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

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

William Yang Wang. "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, 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-2067. URL http://aclweb.org/anthology/P17-2067.

Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni. Generalizing from a Few Examples: A Survey on Few-shot Learning. *ACM Computing Surveys*, 53 (3):63:1–63:34, June 2020. ISSN 0360-0300. doi: 10.1145/3386252. URL https://doi.org/10.1145/3386252.

Zeerak Waseem and Dirk Hovy. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California, 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-2013. URL http://aclweb.org/anthology/N16-2013.

Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov. Demoting Racial Bias in Hate Speech Detection, May 2020. URL http://arxiv.org/abs/2005.12246. arXiv:2005.12246 [cs].

Rahul Yedida and Snehanshu Saha. A novel adaptive learning rate scheduler for deep neural networks. 2019. doi: 10.13140/RG.2.2.28333.95201. URL http://rgdoi.net/10.13140/RG.2.2.28333.95201. Publisher: Unpublished.

Anam Zahid, Maham Nasir Khan, Ahmer Latif Khan, Faisal Kamiran, and Bilal Nasir. Modeling, Quantifying and Visualizing Media Bias on Twitter. *IEEE Access*, 8:81812–81821, 2020. ISSN 2169-3536. doi: 10.1109/ACCESS.2020.2990800. URL https://ieeexplore.ieee.org/document/9079528/.

John R. Zaller. *The Nature and Origins of Mass Opinion*. Cambridge University Press, 1 edition, August 1992. ISBN 978-0-521-40449-5 978-0-511-81869-1 978-0-521-40786-1. doi: 10.1017/CBO9780511818691. URL https://www.cambridge.org/core/product/identifier/9780511818691/type/book.

Savvas Zannettou, Mai Elsherief, Elizabeth Belding, Shirin Nilizadeh, and Gianluca Stringhini. Measuring and Characterizing Hate Speech on News Websites. In *12th ACM Conference on Web Science*, WebSci '20, pages 125–134, New York, NY, USA, July 2020. Association for Computing Machinery. ISBN 978-1-4503-7989-2. doi: 10.1145/3394231.3397902. URL https://doi.org/10.1145/3394231.3397902.

Wangchunshu Zhou, Canwen Xu, Tao Ge, Julian McAuley, Ke Xu, and Furu Wei. BERT Loses Patience: Fast and Robust Inference with Early Exit. In *Advances in Neural Information Processing Systems*, volume 33, pages 18330–18341. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/d4dd111a4fd973394238aca5c05bebe3-Abstract.html.

Arkaitz Zubiaga, Maria Liakata, and Rob Procter. Exploiting Context for Rumour Detection in Social Media. In Giovanni Luca Ciampaglia, Afra Mashhadi, and Taha Yasseri, editors, *Social Informatics*, volume 10539, pages 109–123. Springer International Publishing, Cham, 2017. ISBN 978-3-319-67216-8 978-3-319-67217-5. doi: 10.1007/978-3-319-67217-5_8. URL http://link.springer.com/10.1007/978-3-319-67217-5_8. Series Title: Lecture Notes in Computer Science.

CHAPTER A

Appendix

A.1 Huggingface Models

Table 9. Huggingface models used

Bart	“facebook/bart-base”
Roberta-Twitter	“cardi□nlp/twitter-roberta-base”
Electra	“google/electra-base-discriminator”
GPT-2	“gpt2”
ConvBERT	“YituTech/conv-bert-base”

A.2 Proxy Task Results

Table 10. Proxy task average F_1-Scores

Rank	Model	F_1-Score
1	BART	0.81146
2	RoBERTa-Twitter	0.80851
3	ELECTRA	0.80646
4	GPT-2	0.80371
5	ConvBERT	0.80321
6	RoBERTa	0.80278
7	XLM-ProphetNet	0.80275
8	ERNIE 2.0	0.80138
9	XLNet	0.80056
10	T5	0.79113
11	PEGASUS	0.78841
12	BERT	0.78465
13	SimCSE	0.78345
14	Bloom	0.77225
15	XLM-Roberta	0.77163
16	Mirror-BERT	0.76812
17	ALBERT	0.76396
18	DialoGPT	0.76251
19	Transformer-XL	0.75957
20	infoXLM	0.71616
21	CANINE	0.70034
22	DeBERTa	0.69493
23	Funnel Transformer	0.66016
24	ProphetNet	0.59225
25	CharBERT	0.58124
26	BigBird	0.55395
27	LayoutLM	0.5289
28	Longformer	0.40186
29	XLM-RoBERTa-Twitter	0.38973
30	CharacterBERT	0.26628

Declaration

1. I hereby declare that this thesis entitled:

   Improving Media Bias Detection with state-of-the-art Transformers

   is a result of my own work and that no other than the indicated aids have been used for its completion. Material borrowed directly or indirectly from the works of others is indicated in each individual case by acknowledgement of the source and also the secondary literature used.

   This work has not previously been submitted to any other examining authority and has not yet been published.

2. After completion of the examining process, this work will be given to the library of the University of Konstanz, where it will be accessible to the public for viewing and borrowing. As author of this work, I agree / ~~do not agree~~ [*)] to this procedure.

Konstanz, 14.02.2023 (Date) _____ (Signature)

Erklärung

1. Ich versichere hiermit, dass ich die vorliegende Arbeit mit dem Thema:

   Improving Media Bias Detection with state-of-the-art Transformers

   selbständig verfasst und keine anderen Hilfsmittel als die angegebenen benutzt habe. Die Stellen, die anderen Werken dem Wortlaut oder dem Sinne nach entnommen sind, habe ich in jedem einzelnen Falle durch Angaben der Quelle, auch der benutzten Sekundärliteratur, als Entlehnung kenntlich gemacht.

   Die Arbeit wurde bisher keiner anderen Prüfungsbehörde vorgelegt und auch noch nicht veröffentlicht.

2. Diese Arbeit wird nach Abschluss des Prüfungsverfahrens der Universitätsbibliothek Konstanz übergeben und ist durch Einsicht und Ausleihe somit der Öffentlichkeit zugänglich. Als Urheber der anliegenden Arbeit stimme ich diesem Verfahren zu / ~~nicht zu~~ [*)].

Konstanz, den 14.02.2023 (Unterschrift) _____

[*)] Please delete as applicable / Nichtzutreffendes bitte streichen.