The effects of training with human- vs. machine-labeled news content on media bias perception

Submitted by: David Krieger (01/956788)

Universität Konstanz

Department of Psychology

Social Psychology and Decision Sciences

1. assessor: Dr. Helge Giese

2. assessor: Prof. Dr. Wolfgang Gaissmaier

Konstanz 2022

Abstract

Media bias is a multi-faceted construct influencing individual behavior and collective decision-making. News consumers and editors might benefit from tools reliably marking biased language to mitigate the negative effects of slanted news coverage. To date, media bias research lacks systematic approaches investigating the effects of visual aids on consumers' media bias awareness. Our study represents a first approach combining automatic bias detection methods from Computer Science with psychological research on teaching effects of bias visualizations. We developed *BiasRoBERTa*, a Deep Learning-based language model detecting biased language on sentence level. Our model achieves a state-of-the-art result of 0.814 $F1$ score on an exhaustive bias corpus named BABE (Spinde, Plank, et al., 2021). We collected a representative set of news sentences from BABE and let the model assign bias labels. Based on human-generated labels from BABE and the machine labels, we created simple visualizations highlighting sentence-level bias. In an online survey, we assigned 512 participants to three groups - two intervention groups receiving bias visualizations based on machine labels or human labels and one control group. Our visualizations' bias teaching effects were measured by assessing the participants' bias perception for every sentence. We observed a significant teaching effect for the human-labeled sentences ($d = 0.29$, $p < .05$). Our machine-based bias visualizations did not foster media bias awareness significantly ($d = 0.23$, $p = 0.12$). Our findings indicate that simple visualizations generated by humans increase media bias awareness. However, further research on the automatic detection of biased language is necessary.

Media Bias ist ein vielschichtiges Konstrukt, welches das individuelle Verhalten und die kollektive Entscheidungsfindung beeinflusst. Nachrichtenkonsumenten und Redakteure könnten von Instrumenten zur zuverlässigen Kennzeichnung verzerrter Sprache profitieren, um die negativen Auswirkungen tendenziöser Nachrichtenberichterstattung abzuschwächen. Bislang gibt es in der Media Bias Forschung keine systematischen Ansätze, die die

Auswirkungen visueller Hilfsmittel auf das Bewusstsein der Verbraucher für Medienvoreingenommenheit untersuchen. Unsere Studie stellt einen ersten Ansatz dar, der automatische Methoden zur Erkennung von Media Bias aus der Informatik mit psychologischer Forschung über die Lerneffekte von Bias Visualisierungen kombiniert. Wir entwickelten BiasRoBERTa, ein auf Deep Learning basierendes Sprachmodell, das voreingenommene Sprache auf Satzebene erkennt. Unser Modell erreicht ein State-of-the-Art-Ergebnis von 0.814 $F1$ Score auf einem umfassenden Bias-Korpus namens BABE (Spinde, Plank, et al., 2021). Wir sammelten eine repräsentative Menge von Nachrichtensätzen aus BABE und ließen das Modell Bias-Labels zuweisen. Auf der Grundlage der von Menschen erstellten Labels von BABE und der maschinellen Labels wurden Visualisierungen erstellt, die die Verzerrungen auf Satzebene hervorheben. In einer Online-Umfrage wurden 512 Teilnehmer in drei Gruppen eingeteilt - zwei Interventionsgruppen, die Visualisierungen von Verzerrungen auf der Grundlage von maschinellen oder menschlichen Labels erhielten, und eine Kontrollgruppe. Der Lerneffekt unserer Visualisierungen wurde gemessen, indem wir die Wahrnehmung der Verzerrungen der Teilnehmer für jeden Satz erfragten. Wir beobachteten einen signifikanten Lerneffekt für die von Menschen markierten Sätze ($d = 0,29$, $p < .05$). Unsere maschinenbasierten Visualisierungen förderten das Bewusstsein für Media Bias nicht signifikant ($d = 0,23$, $p = 0,12$). Unsere Ergebnisse deuten darauf hin, dass einfache, von Menschen erstellte Visualisierungen das Bewusstsein für Medienvoreingenommenheit erhöhen. Weitere Forschungen zur automatischen Erkennung von voreingenommener Sprache sind jedoch notwendig.

Keywords: Media bias, media bias perception, automatic media bias detection

The effects of training with human- vs. machine-labeled news content on media bias perception

Contents

The effects of training with human- vs. machine-labeled news content on media bias perception

An increasing number of consumers nowadays have access to the world wide web, and the primary way of reading news is via online platforms replacing traditionally printed formats continually (Dallmann et al., 2015; Houston et al., 2011; Kaye & Johnson, 2016). Online news provide information from diverse sources based on a wide range of perspectives leading to a higher degree of self-determination in how people gather knowledge (Hamborg et al., 2018). Unfortunately, the diversity of online news opens the door for slanted and non-neutral news coverage. Biased news coverage - referred to as *media bias* in the literature - occurs if subjective reporting on a specific event replaces objective coverage. Media bias manifests in various forms such as bias by word choice or bias by omission of information (Alonso et al., 2017; Hamborg et al., 2018).

This thesis represents an interdisciplinary approach investigating the teaching effects of visualizing human- and machine-labeled news content on the perception of media bias. We combine techniques from Computer Science to detect media bias on a fine-grained linguistic level automatically with methods from Psychology to examine how machine-labeled and human-labeled bias instances are perceived by humans.

Forms of media bias

Media bias is a multi-faceted construct and occurs in numerous forms. The concrete definition of slanted news reporting depends on the underlying research scope of media bias studies, which is one of the fundamental obstacles in media bias research (Baron, 2006; Druckman & Parkin, 2005b; Hamborg et al., 2018).

In many cases, news bias is the result of various interest groups being involved in different stages of the news production and consumption process. Owners, editors, and journalists induce diverse biases when producing news, while cognitive biases occur on the side of consumers (Baron, 2006).

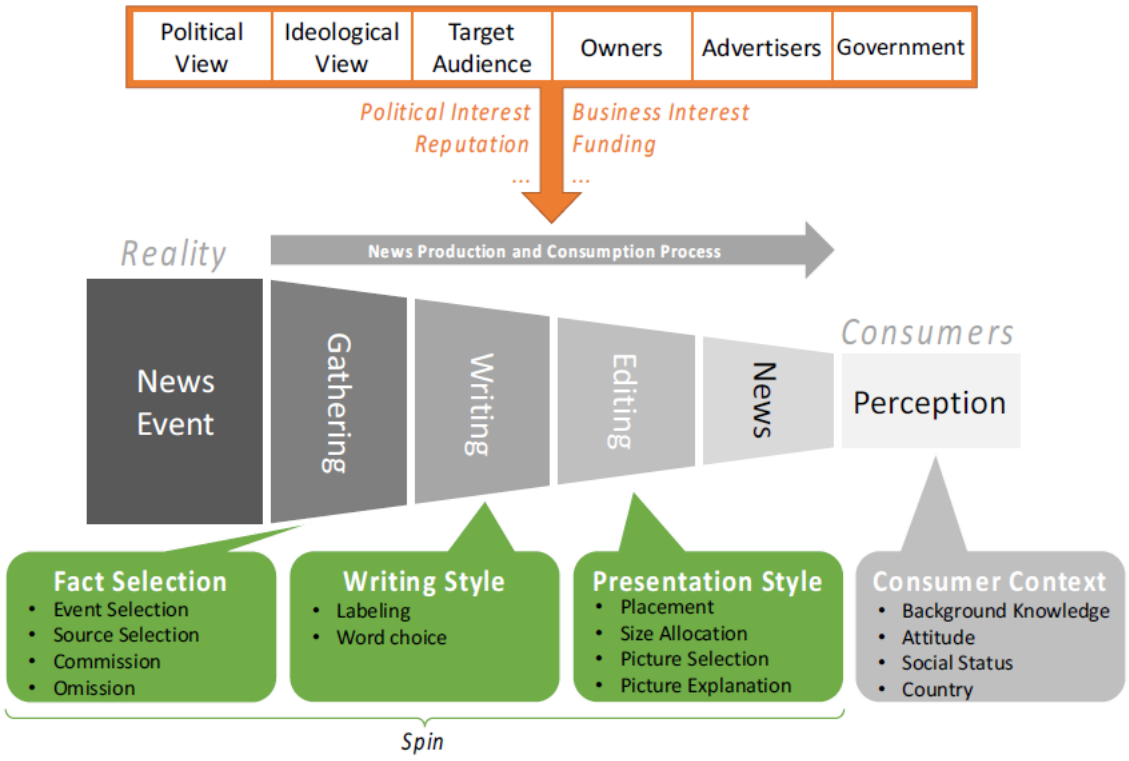
According to Mullainathan and Shleifer (2002), *ideological bias* occurs once

reporters embed particular world views or ideologies in their news coverage. Baron (2006) introduces *partisan bias* as news reporting which supports particular policies introduced by political parties.

In their literature review on media bias, Hamborg et al. (2018) introduce a conceptualization modeling forms of biases evolving in different stages of the news production and production process (see Figure 1).

Figure 1

Forms of media bias occurring in different stages of the news production and consumption process and induced by various interest groups (Hamborg et al., 2018)



Their approach distinguishes between the *Gathering or fact selection stage*, the *Writing stage*, and the *Editing stage*: in the gathering stage, the news production process begins with selecting events to be reported on and gathering additional sources to rely on when reporting. In this stage, *bias by omission* (Alonso et al., 2017) might occur due to the omission of newsworthy information. In the writing stage, journalists put their ideas

and reports on paper. Individual writing styles might lead to linguistic biases referred to as *bias by word choice and labeling*, which is the bias whose perception we examine in the present thesis. In the editing stage, editors use different styles to present news to their consumers. News coverage can vary in placement and size allocation to set the focus on particular events, and pictures can be selected to convey emotions.

A similar conceptualization of bias forms distinguishes between gatekeeping, coverage, and statement bias (D'Alessio & Allen, 2000; Kaye & Johnson, 2016) strongly resembling the above definitions. Gatekeeping bias can be considered an equivalent to ideological bias occurring in the gathering stage, coverage bias refers to the visibility of particular events (gathering stage), and statement bias describes unfair news reporting (writing stage).

The most important bias form for our work is *perceptual bias* (Kaye & Johnson, 2016). It is the only bias form that is entirely isolated from the news production process. Perceptual bias occurs once news consumers perceive news coverage as going against their world views and interests.

Effects of media bias

Media bias can be induced intentionally, but also in a subtle manner without any awareness on the reporter's side. Either way, slanted news coverage has the potential to influence an individual's opinion, its political activity, and media usage (Ardèvol-Abreu & Gil de Zúñiga, 2017; Kaye & Johnson, 2016; Rojas, 2010). On a higher level, it also alters collective decision-making (Aggarwal et al., 2020; Bernhardt et al., 2006; Mitchell, 2014).

When people perceive news as biased, they tend to increase political participation by dealing with online and offline debates on political issues (Rojas, 2010; Weeks et al., 2017). As underlying mechanism motivating growing engagement, Rojas (2010) indicates the willingness to correct what was previously perceived as biased.

Perceived media bias also influences media usage. According to a survey conducted by Kaye and Johnson (2016) during the 2012 presidential election in the U.S., participants

perceiving news coverage as being anti-Obama spent less time reading conservative news. Vice versa, individuals perceiving the media as being biased against the conservative candidate Mitt Romney avoided consuming liberal news sources. A survey by Ardèvol-Abreu and Gil de Zúñiga (2017) shows that the evidence regarding the effects of media bias on media usage is not clear. In contrast to Kaye and Johnson (2016)'s findings, the authors report that participants perceiving news as biased spent less time reading news in general rather than switching to attitude-congruent news.

Investigating the effects of media bias on collective decision-making, the literature provides some evidence that media bias and voting behavior are linked (Bernhardt et al., 2006; Mitchell, 2014). Bernhardt et al. (2006) state that ideological polarization represents a substantial factor influencing election outcomes. Partisans of rival and opposing parties tend to share their opinion merely with people having similar political views. Beyond that, they mostly read news matching with their own beliefs (Mitchell, 2014). From a psychological point of view, such information cocoons can be considered a form of *filter bubbles* and *echo chambers* emerging due to *selective exposure*, among others (Garrett, 2009; Sindermann et al., 2020; Spohr, 2017)[1]. According to Bernhardt et al. (2006), ideological polarization acts as a mediator between media bias and voting behavior. News outlets customize their coverage based on consumers' opinions and interests to maximize profit which in turn leads to a distorted presentation of information and an unbalanced distribution of voter ideologies. A further crucial point is that voters on both ends of the political spectrum have the highest impact on election results since they show the highest participation rates in elections and political activities (Mitchell, 2014).

Bernhardt et al. (2006)'s statements regarding the link between slanted news coverage and voting decisions are supported by a data-driven approach from Druckman and Parkin (2005a) investigating how editorial slant shapes voting behavior by combining

---

[1] In the remainder of this work, we discuss in more detail how the phenomenon of media bias is linked with concepts and theories from Psychology.

exhaustive content analysis of two editorially distinct newspapers with an election day exit poll. The authors report a direct association between biased news reporting and voting outcomes.

How is media bias rooted in Computer Science?

The present thesis compares human perceptions of machine-labeled and human-labeled news content. Nowadays, sophisticated information techniques to label text data automatically exist. Thus, we want to outline how media bias research is linked to Computer Science. In the following, we briefly elaborate on important key concepts from subfields of Computer Science that are crucial to getting high-level insights on how Computer Scientists develop algorithms and language models that are useful for media bias research. Beyond that, we present existing approaches on the creation of bias corpora, automized detection of media bias, and how recent developments are incorporated in the present thesis.

Computers Scientists have conducted research on media bias for some years. Most of the media bias studies published in the Computer Science domain deal with the automatic detection of media bias on different linguistic levels.

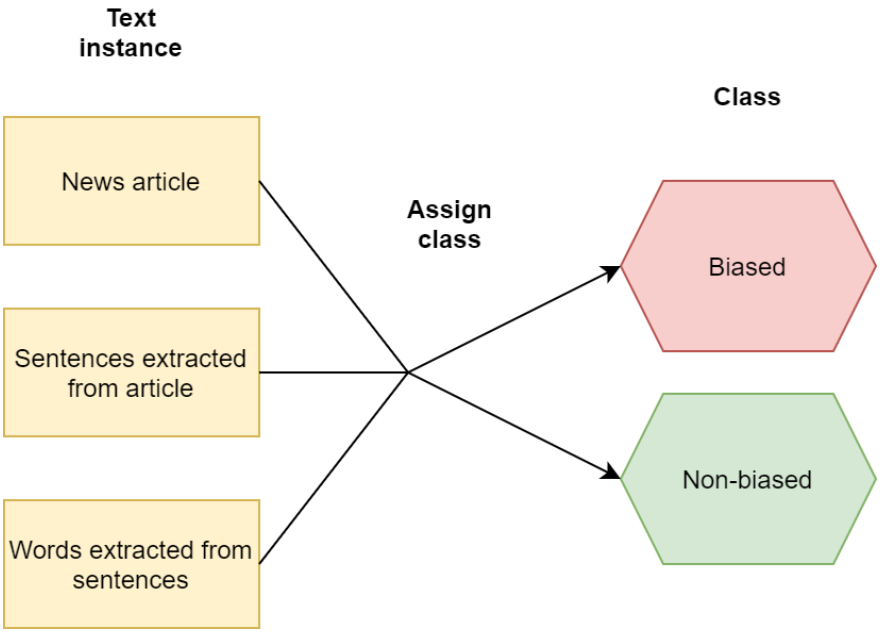
When performed by human beings, detecting media bias instances in news through qualitative content analysis is a laborious and resource-consuming task. In Computer Science, the overall goal is to hand the job of detecting media bias over to a machine. Considering the sheer amount of news published daily on web platforms, automizing the task of identifying media bias in news corpora represents a scalable solution (Hamborg et al., 2018). Therefore, combining information technologies with findings in Social Sciences is of utmost importance to promote media bias detection.

Mostly, existing studies use techniques from *Natural Language Processing* (NLP) to analyze news. NLP is a subfield of Computer Science and *Artificial Intelligence* dealing with processing and analyzing natural language data, i.e. written and spoken data generated by humans (Liddy, 2001). Nowadays, NLP techniques are implemented in a wide

range of applications such as e-mail filters (filtering spam mails), smart assistants (Amazon Alexa), search engines (Google), language translation, or text analytics[2]. When using NLP techniques to detect media bias, a *text classification* task is to be solved. Text classification aims at assigning given text sequences to pre-defined classes/categories. In the context of media bias research, text classification can be performed, among others, on word level (Recasens & Jurafsky, 2013; Spinde, Rudnitckaia, Mitrovic, et al., 2021), sentence level (Hube & Fetahu, 2018; Spinde, Plank, et al., 2021), or article level (Chen et al., 2020) meaning that a text instance (word, sentence, or article) is assigned to a class (e.g biased vs. non-biased) (see Figure 2).

Figure 2

Media bias detection formulated as a text classification task



Within artificial intelligence, there are two fundamental approaches on how algorithms learn to find patterns in existing data: *Supervised Learning* and *Unsupervised Learning*. Text classification is a Supervised Learning task requiring *labeled data*. Supervised Learning algorithms use existing output labels to model the relationship

[2] https://www.tableau.com/learn/articles/natural-language-processing-examples

between input (text) and output (assigned class/category) (Cunningham et al., 2008). In contrast, Unsupervised Learning uses *unlabeled data* to find hidden patterns in the data that can be used to cluster data into categories (Hastie et al., 2009). Thus, Supervised Learning presupposes pre-defined output categories/classes whereas unsupervised learning aims at structuring the data into categories/classes.

Implementing algorithms aiming to model biased language requires conceptual and manual groundwork: first, linguists and psychologists have to come up with a clear definition of media bias. Second, bias detection algorithms need a certain amount of labeled data to learn from. Prior to implementing and training NLP models to identify slanted reporting, humans are required to manually annotate news data in terms of bias.

Varying media bias perception in crowdsourcing studies

Several approaches in Computer Science have tackled the problem of creating gold-standard data sets representing media bias exhaustively (Färber et al., 2020; Hube & Fetahu, 2018; Lim et al., 2020; Recasens & Jurafsky, 2013; Spinde, Plank, et al., 2021; Spinde, Rudnitckaia, Kanishka, et al., 2021). Mostly, existing studies rely on crowdsourcing to collect labeled bias data. However, one of the major shortcomings is that clickworkers show variations in terms of media bias perception. There are no crowdsource-based studies reporting at least fair *interrater agreement* (IRA). Lim et al. (2020) create a data set containing bias instances on word and sentence level with IRA scores ranging from Fleiss' κ = -0.0824 to 0.0004. Färber et al. (2020) gather 2000 sentences on which clickworkers show an agreement of Krippendorff's α = -0.05. To the best of our knowledge, the data set created by Spinde, Rudnitckaia, Kanishka, et al. (2021) represents the most exhaustive bias datasets collected through crowdsourcing achieving an IRA of Fleiss' κ = 0.21. Based on their findings, the authors assume that crowdsourcers might not be able to render accurate bias labels resulting in low agreement scores. Spinde, Rudnitckaia, Kanishka, et al. (2021) argue that bias might occur in subtle forms requiring domain knowledge and expertise in terms of its identification. Therefore, the authors

perform a second study in which they ask media bias experts to label the data used in their first approach (Spinde, Plank, et al., 2021). Relying on experts results in a substantial IRA increase to Krippendorff's $\alpha = 0.39$ (fair agreement according to Krippendorff (2018)). Based on the presented results, we can assume that in the broader public (represented through crowdsourcers here), bias perception varies strongly across individuals.

Existing bias detection approaches

Several approaches have tackled the problem of identifying media bias automatically. Existing detection techniques vary substantially, and most approaches refer to encyclopedia data to train their models. In the following I briefly present key papers dealing with media bias detection.

Recasens and Jurafsky (2013) implement a method to detect bias-inducing words in a given biased sentence. Their approach is based on Wikipedia data. The authors collect articles going against Wikipedia's *Neutral Point of View*[3] and extract sentences and their revised versions assuming that unrevised sentences contain biased words. Thereupon, they define linguistic cues inducing biased language and assign them to words. The features are then passed to a logistic regression model predicting a bias label for every word. The model achieves 34% accuracy for predicting the word with the highest probability of being biased within a sentence. Although the model does not show sufficient performance, the authors point out that crowdsourcers merely achieve 37% accuracy on the same data. Thus, bias detection on word level seems to be challenging for both humans and machines.

Hube and Fetahu (2018) tackle the bias detection problem on sentence level relying on similar data as Recasens and Jurafsky (2013). However, their modeling approach differs strongly: the authors implement a *neural network* architecture generating word representations in form of vectors. The language representations encode semantic and syntactic information about every word that can be used, for example, in text classification

---

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

tasks. The best-performing model achieves 82% accuracy and 0.77 $F1$ score %[4].

To the best of our knowledge, the best-performing approach with the highest relevance for bias detection in news is published by Spinde, Plank, et al. (2021). The authors implement a neural net-based model detecting bias on sentence level which is trained on a state-of-the-art media bias dataset (BABE = Bias Annotated By Experts) containing an exhaustive amount of bias instances from multiple news outlets covering various topics. Their best-performing model achieves 0.804 $F1$ score.

Using bias detection approaches to teach media bias in a scalable manner

In this thesis, we make use of recent developments in Computer Science to detect media bias on sentence level. We implement a neural net-based Deep Learning model based on top of Spinde, Plank, et al. (2021)'s approach. Then, we let the model assign bias labels to a set of sentences extracted from BABE (Spinde, Plank, et al., 2021). Biased sentences are visually highlighted and presented to participants of our media bias perception survey within a training phase. In a subsequent test phase, participants are then asked to annotate sentences themselves in terms of existing media bias. Thus, one of our goals is to make use of automized bias labeling approaches to teach media bias in a scalable manner.

How is media bias rooted in Psychology?

The phenomenon of media bias is linked with several key concepts emerging from research in Social Psychology. Two of these key concepts are *Cognitive Dissonance* (Festinger, 1957) resulting in selective exposure (Klapper, 1960).

In his book *An Introduction to the Theory of Cognitive Dissonance*, Festinger (1957) states that "[...] the individual strives towards consistency within himself." (p.1). Once personal attitudes and actual behavior conflict with each other, and an individual cannot explain this inconsistency, psychological discomfort occurs. In this context, Festinger introduces the notion of *dissonance* motivating the person to reduce psychological

[4] In the methods section, we briefly introduce the $F1$ score since it is an important metric to evaluate classification algorithms

discomfort. A person who is smoking knowing about respective health consequences might experience *cognitive dissonance* leading to either a reduction of the dissonance-inducing behavior (quit smoking) or the avoidance of situations/information increasing the dissonance (e.g. discussions with friends about the health consequences of smoking).

Selective exposure can be considered a consequence of cognitive dissonance (Tsang, 2017). More specifically, selective exposure is a "[...] process through which people avoid or reduce cognitive dissonance." (Williams et al., 2016). In his study on the effects of mass communication, Klapper (1960) introduces the selective exposure theory referring to an individual's tendency to search for information that is congruent with individual attitudes and beliefs.

Transferring the concept of selective exposure to news consumption, the theory states that individuals merely read news spreading information and opinions that are congruent with own beliefs. If this is the case, news consumers risk being caught up in filter bubbles and echo chambers (Garrett, 2009; Sindermann et al., 2020; Spohr, 2017).

How visual aids foster media bias awareness: a psychological learning perspective

Our study uses bias visualizations to teach participants how slanted news coverage manifests linguistically. We assume that subjects learn how to detect media bias on sentence-level in the training phase and transfer acquired bias knowledge to the test phase. From a psychological learning perspective, the underlying learning mechanism can be considered a form of *Associative Learning*.

According to the Associative Learning theory, a person learns a relationship between two stimuli by association. The link between the stimuli can be generated through repeated pairing of the stimuli, also known as *Classical Conditioning* (Siegel, 1983). The conditioning effect can be measured by an initially unconditioned response (UR), which is evoked by an unconditioned stimulus (US) at the beginning. Through repeated pairing of the US and another initially neutral stimulus, the response is conditioned on the neutral

stimulus, which then becomes the conditioned stimulus (CS). Once the response is conditioned on the CS, the response can be evoked through exposure to the CS only (without exposure to the US).

A systematic approach showing the mechanisms of Classical Conditioning in terms of animal behavior was conducted by Ivan Pavlov (Pavlov, 1949). In his first experimental trials, Pavlov fed his dogs with meat (unconditioned stimulus), and the animals started to salivate (unconditioned response). In the subsequent trials, Pavlov rang a bell (neutral stimulus) every time the dogs were fed. After repeated pairing of this neutral stimulus and the US, the dogs salivated once the bell rang. Consequently, the animals' response was conditioned on the initially neutral stimulus.

Connecting the insights on Associate Learning with our study, we state that our visual bias-highlighting aids teach participants to label sentences in terms of bias without being exposed to bias visualizations, in the test phase. Through repeated pairing of sentences and respective bias highlighting, we assume that participants learn a relationship between the two stimuli, empowering them to provide accurate bias labels in the test phase. Consequently, our bias visualizations can be considered a key component to foster media bias awareness.

How to increase media bias awareness

Communicating media bias effectively to the broader public is crucial to foster reflective news reading and increase *media bias awareness* (Spinde, Jeggle, et al., 2021). However, media bias research is still in its infancy, and most approaches dealing with media bias rather focus on conceptual work and detection techniques than on communicating biased news content.

Cook et al. (2017) investigate how misinformation about climate change can be communicated to the broader public. They use inoculation messages to raise awareness for misinformation and report that their intervention neutralizes adverse effects of slanted coverage. Munson and Resnick (2013) develop a browser widget displaying feedback

regarding a user's weekly political leaning of news consumption. Their approach leads to a slightly more balanced reading behavior. Park et al. (2009) introduce *NewsCube*, a news service automatically creating and providing different viewpoints on a news event of interest to mitigate effects of biased news coverage.

To the best of our knowledge, the only study systematically investigating the effects of different bias communication and visualization strategies on media bias awareness is performed by Spinde, Jeggle, et al. (2021). The researchers select three methods to communicate bias in news articles: text annotations (manual highlighting of biased words/phrases), an inoculation message (brief introduction on media bias and how it manifests), and political classifications of news articles. 985 participants were exposed to either a liberal or conservative news article based on the visual aids. Thereupon, subjects were asked to indicate how biased they perceive the presented article. The authors report that highlighting biased words and phrases in news articles (text annotations) and the forewarning message increase media bias awareness significantly, whereas presenting a news article's political category does not foster bias awareness.

Purpose of this thesis

The present thesis is built upon crucial findings from Spinde, Jeggle, et al. (2021). We use their visualization strategy of highlighting biased phrases (text annotations) to increase media bias awareness. However, we do not solely rely on manual bias labels provided by humans. Using manual text annotations does not represent a scalable solution to uncover media bias due to the massive amount of published news content. We implement both a visualization approach based on human labels and machine annotations.

Our survey is structured into two phases: in the first so-called teaching phase, we present highlighted sentences to teach how media bias manifests linguistically in the news. In a subsequent test phase, we ask users to label unannotated sentences in terms of underlying bias themselves. We measure participants' accuracy of labeling biased text instances based on a BABE - a gold-standard media bias data set (Spinde, Plank, et al.,

2021). Thus, we can measure how human-labeled and machine-labeled news content raises media bias awareness.

We assign participants to three groups: two intervention groups, and one control group. In both intervention groups, our media bias teaching consists of the visualizations. We assume that participants learn the relationship between sentences and visualizations through Associative Learning. For the first group, we use human-labeled sentences, and the second group is assigned to the machine-labeled sentences. The third group is a control group for which we do not present any bias visualization/highlighting. However, we can assume that participants assigned to the control group get some insights about media bias through mere exposure to the sentences, which can be considered a form of *Experiental Learning* (Kolb, 2014).

Our main research question is to what magnitude simple visual aids highlighting biased sentences foster media bias awareness. Also, we want to specifically address the question concerning the benefit of machine-generated bias labels. More information about the survey and the Deep Learning algorithm identifying biased news coverage on sentence-level is provided in the remainder of this work.

Research objectives

Based on the introduced findings regarding the detection and visualization of biased news coverage and the presented motivational background, we aim to achieve the following objectives.

1. Implement a Deep Learning algorithm to classify biased news coverage on sentence level.
2. Extract a representative sample[5] of sentences from Spinde, Plank, et al. (2021) and pass the sentences to the implemented model to obtain binary bias labels (*Biased* vs. *Non-biased*). Human-labeled sentences are extracted from BABE (Spinde, Plank,

[5] We define representativeness as a balanced distribution of news articles from which sentences are extracted in terms of political orientation.

et al., 2021).

3. Create simple bias visualizations highlighting sentences based on the obtained labels.
4. Perform an online survey to investigate the visualizations' effects on media bias awareness.

Hypotheses

In the light of visualization effects on media bias awareness reported by Spinde, Jeggle, et al. (2021), we make the following assumptions.

H1 (directional): Both intervention groups achieve a significantly higher accuracy score in rating sentences in terms of bias than the control group in both the teaching and test phase.

H2 (non-directional): Both intervention groups do not differ significantly regarding their labeling accuracy in both the training and test phase.

H3 (directional): Since we do not provide any bias highlighting in the test phase, we assume an interaction effect between the visualizations and the survey phases on the participant's accuracy of rating the sentence in terms of existing bias. In the test phase, the accuracy of the intervention groups should be higher than in the control group. However, we assume that the effect is smaller than in the teaching phase.

Methods

In this section, we take a detailed look at tools and methods incorporated into our media bias study. First, we elaborate on the collection of our sentence survey sample and the development of our automized sentence-level bias classifier as well as the resulting sentence-level bias visualizations. Then, we describe our online survey and respective measures.

Collection of sentence sample

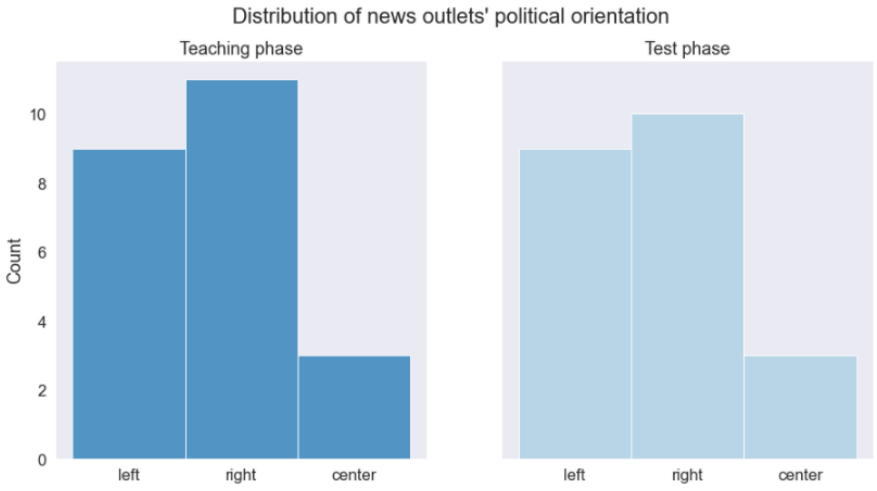
The first preparatory step for our media bias perception study comprises the collection of biased and non-biased news sentences building the data ground-truth for our subsequent online survey: we extract sentences from the state-of-the-art media bias data

set BABE, published by Spinde, Plank, et al. (2021). Initially, the corpus contains 3700 sentences extracted from diverse US news outlets covering a wide range of topics. The data also includes binary bias labels (*Biased* vs. *Non-biased*) assigned by five media bias experts[6].

For the survey, we extract 46 sentences from BABE intending to collect a representative sentence sample in terms of the underlying political distribution of news outlets. Figure 3 and 4 show the overall distribution of news outlets and their respective political leaning from which the sentences were extracted.

Figure 3

Political distribution of news outlets according to allsides.com



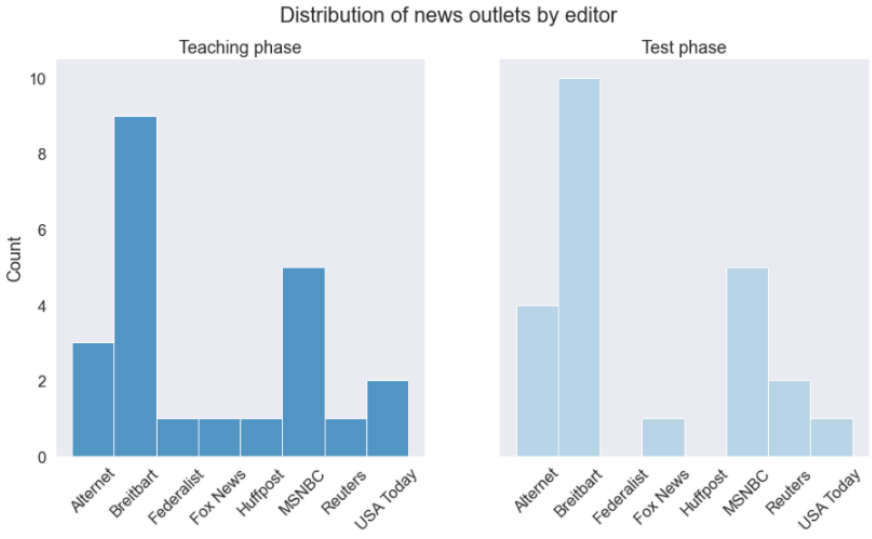
Human- and machine-generated bias highlighting of news sentences

The present thesis uses simple visualizations of slanted language to teach how media bias manifests in news coverage. To visualize biased sentences, we first need to come up with labels indicating if a sentence contains biased language or not. Here, we make use of two different approaches: first, we used human-generated binary bias labels from BABE (Spinde, Plank, et al., 2021). Second, we exploit current research findings on the automated detection of biased news coverage and implement a state-of-the-art neural-based

[6] Experts defined as individuals working in the context of media bias.

Figure 4

Distribution of news outlets by editor



classification model categorizing sentences as *Biased* and *Non-biased*.

Due to the massive amount of daily published news, scalable bias detection methods need to rely on automized text processing algorithms from Computer Science. Thus, we compare both the effects of human- and machine-generated bias labeling approaches on fostering media bias awareness. In the following, we introduce the architecture and training process of our state-of-the-art language model detecting media bias on sentence level.

Bias classifier

Sophisticated information techniques exist to analyze text data automatically. Our aim of detecting bias on sentence level can be considered a form of text classification using techniques from Natural Language Processing. In general, we want to hand over the job of labeling news sentences in terms of existing bias (*Biased* vs. *Non-biased*) to an algorithm. Here, we used a supervised learning approach searching for the relationship between input data (sentences) and output labels (binary bias labels). In the future, we are hopefully able to use the trained algorithm to detect bias in a large number of unlabeled sentences, that is to say, raw news articles.

Similar to the state-of-the-art bias detection approach by Spinde, Plank, et al.

(2021), we use a language model called *RoBERTa* and train it on a large corpus containing biased and neutral text sequences (Pryzant et al., 2020).

Advantages of transformer-based text processing

Mostly, automized bias detection approaches are based on bias-inducing linguistic features that are passed to a classifier using the features to distinguish biased from non-biased news content (Hube & Fetahu, 2018; Recasens & Jurafsky, 2013; Spinde, Rudnitckaia, Mitrovic, et al., 2021).

Although feature-based detection approaches partly achieve acceptable results, the linguistic subtleness of slanted news coverage is emphasized to be a great challenge for automized classification methods (Spinde, Plank, et al., 2021). Subtle differences such as "illegal immigrants" vs. "undocumented immigrants" or "climate change" vs. "global warming" (Schuldt et al., 2011) are challenging to identify for existing computational methods. Thus, we need language models capable of acquiring a deep semantical and syntactical understanding of words and phrases.

The *Transformer* architecture (Vaswani et al., 2017) can process words based on their surrounding context. Since the semantic and syntactic meaning of a word is substantially influenced by its context (Langendoen, 1964), the architecture can represent implicit linguistic features automatically, which are encoded in vectorized language representations - so-called *word embeddings*. The vectors can be used for various NLP tasks requiring semantic/syntactic understanding, among others. In contrast to contextualized word embeddings by other architectures such as *Recurrent Neural Networks*, Transformers process sequential data in a parallel manner avoiding the *vanishing gradient problem* (Hochreiter, 1998; Wolf et al., 2020).

Domain-adaptive pre-training with RoBERTa

Many state-of-the-art NLP models such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and the GPT models (Budzianowski & Vulić, 2019) are based on the Transformer architecture trained on large amounts of text data. The models yield word

embeddings encoding general language information that can be fine-tuned on various downstream tasks such as text classification.

To the best of our knowledge, only a few approaches incorporate state-of-the-art NLP models into research on media bias detection (Spinde et al., 2022; Spinde, Plank, et al., 2021). The main challenge for such approaches is to create massive datasets containing biased and non-biased text (Spinde, Plank, et al., 2021; Spinde, Rudnitckaia, Kanishka, et al., 2021). Equipping a large-scaled language model[7] with an understanding for biased language requires respective data to pre-train and fine-tune the model.

The pre-training of our state-of-the-art bias detection model is oriented on two recently published approaches on media bias detection using transformer-based models pre-trained on bias-related data:

Spinde, Plank, et al. (2021) pre-train various models such as BERT, RoBERTa, and DistilBERT (Sanh et al., 2020) using *distant supervision learning* on news headlines from articles with different political leanings and fine-tune it on BABE. Their best performing models classify biased/non-biased sentences extracted from BABE with 0.804 $F1$ (BERT) and 0.799 $F1$ (RoBERTa), respectively. The distant supervision training improves model performance by up to 1.5% compared to the baseline model.

Another approach by Spinde et al. (2022) trains DistilBERT on various combinations of bias-related datasets using *Multi-task Learning* (MTL). Their best-performing MTL model achieves 0.776 $F1$ score on a subset of BABE, resulting in an improvement of 3.0% compared to the simplest baseline model. However, the MTL model is outperformed by a baseline model trained on a subset of the MTL datasets with an improvement of 3.6% ($F1 = 0.782$). The respective model is trained on a subset of the *Wiki Neutrality Corpus* (Pryzant et al., 2020) containing 180k sentence pairs from English Wikipedia. The sentences are selected from Wikipedia articles going against the platform's *Neutral Point of View* standard. The pairs contain the original and revised version of a

[7] e.g. BERT consists of 110 million parameters (Devlin et al., 2019).

sentence. Spinde et al. argue that the Wiki Neutrality Corpus might be strongly bias-related, helping to model to generate an understanding for biased language.

In our work, we aim to exploit this finding and extend the pre-training of a BERT-like model on the whole Wiki Neutrality corpus of 180k sentence pairs[8]. We expect that training the model on large amounts of bias-related data results in good performances in our sentence-level bias labeling task.

Training the model on bias-related data can be considered as *Domain-adaptive pre-training*. Adapting a pre-trained language model to a specific domain becomes essential when the target domain differs strongly from the pre-training ground truth. The domain of media bias is different from the domain BERT-like models are pre-trained on. For example, BERT is trained on English Wikipedia and the *BooksCorpus* (Zhu et al., 2015) whereas media bias mostly deals with news data. Thus, we need to adapt our model to biased and unbiased news data distribution.

Han and Eisenstein (Han & Eisenstein, 2019) refer to BioBERT (Lee et al., 2020) and SciBERT (Beltagy et al., 2019) as approaches to re-training a BERT-based model on the biomedical and scientific domain, respectively. Sun et al. (2019) explore different techniques for domain adaptive pre-training of BERT for text classification tasks such as sentiment classification, question classification, and topic classification. The transformer is pretrained on data from various domains leading to performance boosts on many tasks if the training data are related to the target task's domain. Pre-training on data from unrelated domains does not improve the performance or even decreases the classification accuracy. The results are supported by Gururangan et al. (2020) investigating domain-adaptive pre-training of RoBERTa in four different target domains (biomedical and computer science publications, news, and reviews) and eight subsequent classification tasks. Their results show that domain-adaptive pre-training consistently improves classification performance in all domains.

---

[8] Compared to only 50k pairs in Spinde's approach.

For our domain-adaptive pre-training approach based on the Wiki Neutrality corpus, we rely on RoBERTa - a robust BERT-like model achieving state-of-the-art results in many NLP tasks. The training process and experiments are described in the following section.

Domain-adaptive experiments

Our domain-adaptive pre-training is based on the Wiki Neutrality Corpus containing 180k sentence pairs labeled as *Biased* and *Non-biased*. We initialize RoBERTa with pre-trained weights provided by the *HuggingFace* API[9], and stack a dropout layer (Dropout = 0.2) and randomly initialized linear transformation layer (768,2) on top of the model.

For the domain-adaptive pre-training, we implement an 80/20 train/test split. Sentences are batched together with 32 sentences per batch. For model optimization, we use the *AdamW* optimizer[10] with a learning rate of $1e^{-}5$, and model performance is evaluated on binary cross-entropy loss. Model convergence can be observed after two epochs and a runtime of $\sim$ 9 hours on a Tesla P100-PCIE GPU with 16GB RAM.

We fine-tune and evaluate the model on BABE (Spinde, Plank, et al., 2021) with a batch size = 32. We again use the AdamW optimizer (learning rate = $4e^{-}5$), and model convergence based on cross-entropy loss can be observed after 3-4 epochs. Due to the small data size of 3700 sentences, we report the model's $F1$ score in the binary bias labeling task averaged by 5-fold cross-validation. Fine-tuning is performed on a Tesla K80 GPU (12GB RAM) taking $\sim$ 20 minutes.

Our domain-adaptive RoBERTa model (*BiasRoBERTa*) achieves **0.814** $F1$ (SD = 0.004) on BABE which can be considered a new and significant[11] state-of-the-art result compared to the previous best-performing model (Spinde, Plank, et al., 2021) measured by a McNemar test statistic ($\chi^2 = 3.84, p = 0.049$).

---

[9] https://huggingface.co/

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

[11] significant on $\alpha = 0.05$

Final bias visualizations

We use BiasRoBERTa to generate binary bias labels (*Biased* vs. *Non-biased*) for our sentence survey sample. The labels are then used for sentence-level bias highlighting in our survey's second intervention group.

For our first intervention group, we rely on human-generated binary bias labels extracted from BABE. An exemplary bias highlighting is shown in Figure 5. Figure 6 shows the class distribution of biased vs. non-biased sentences in both intervention groups.

Figure 5

An example of our sentence-level bias highlighting (above) vs. a non-biased sentence (below).

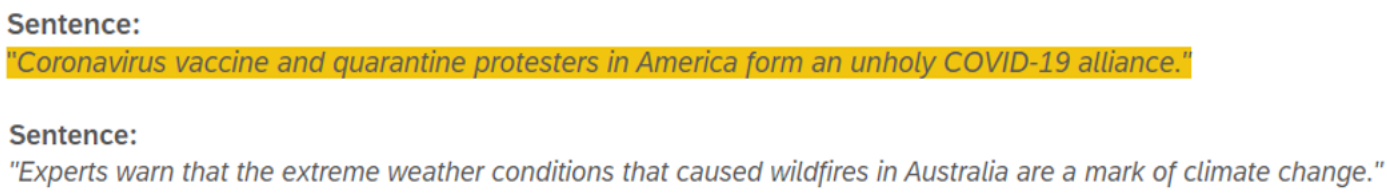
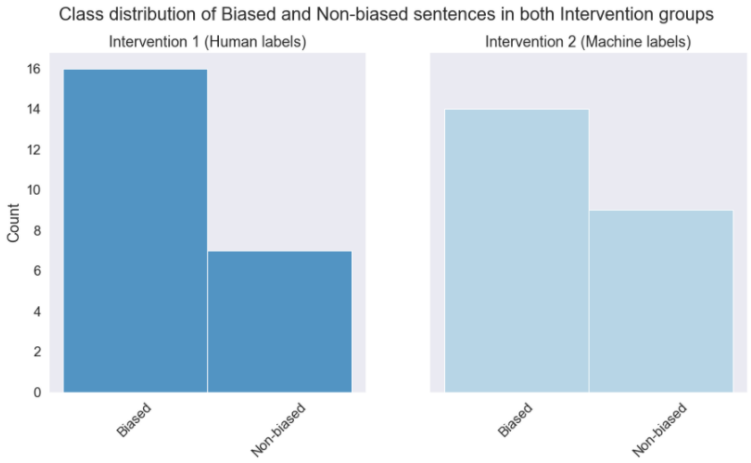


Figure 6

Bias distribution in both intervention groups

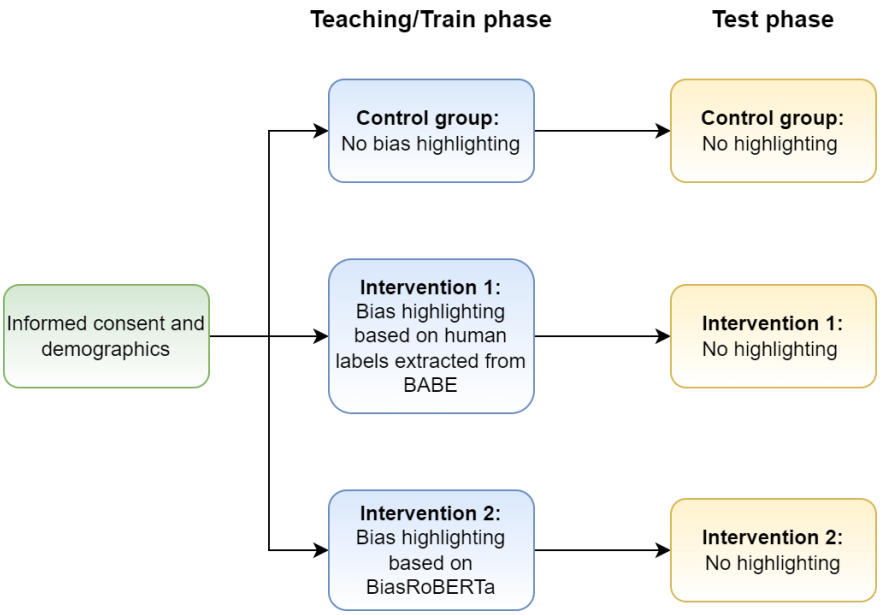


Survey

Our media bias perception survey consisted of a teaching and test phase and was created on *Qualtrics*[12]. We assigned 23 sentences to the teaching and 23 sentences to the test phase. The teaching phase's sentences contained bias visualizations for both intervention groups. The control group did not get any bias highlighting. In the test phase, participants of all groups were exposed to the raw sentences without any highlighting. Figure 7 outlines the survey flow.

Figure 7

Survey flow



First, participants were presented general with information about the study, and demographics such as age, gender, proficiency in English, and political orientation were collected. Participants were randomly assigned to one of the three survey groups. The study started with a short introduction to the concept of media bias and how it manifests linguistically in news reporting. In the subsequent teaching/training phase, we exposed participants to the first 23 sentences with optional bias highlighting in the intervention

[12] https://www.qualtrics.com/de/?rid=langMatch&prevsite=en&newsite=de&geo=DE&geomatch=

groups. We asked them to rate the degree of existing bias. The test phase started with a short message indicating that participants completed the first half of the study. Then, we presented the second set of 23 sentences and again asked for bias ratings. In the last survey part, we asked if we could trust the collected data for scientific research and provided the opportunity to comment on the study.

Data collection

We recruited 512 participants on *Prolific*[13] taking part in the survey. Per completed trial, we paid 1.50£, which is proportional to an hourly salary of 7.73£ (corresponding to an average payment according to Prolific). The survey was estimated to take 12 minutes, and all participants took part voluntarily and gave informed consent.

We used Prolific's feature of assigning a representative sample of sex, age, and ethnicity. The platform uses census data from the US Census Bureau or the UK Office of National Statistics to divide the sample into subgroups with the same proportions as the national population.

To calculate a target sample size, we performed a power analysis with the software program G Power[14]. Our goal was to obtain .80 power to detect a small to medium effect size of $f = 0.15$ at the standard 0.05 α error probability. The software proposed a target sample size of 400-500 participants. The survey was published on 11th November 2021 and data collection was completed within 24 hours.

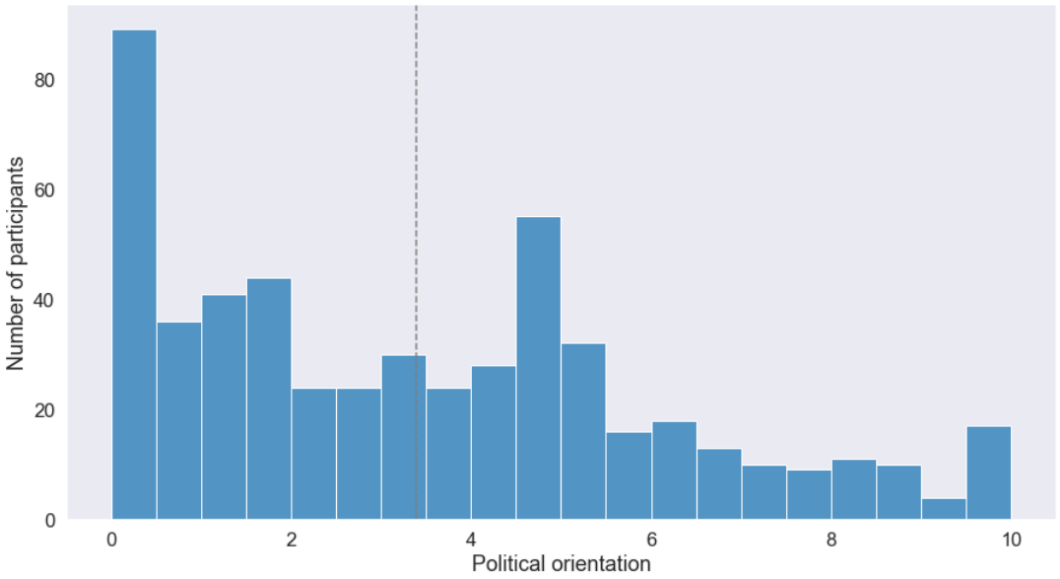
All items were based on a force response format. Thus, missing data did not occur. We merely excluded participants failing one of the attention checks or indicating that we cannot trust their data for scientific research. Based on these exclusion criteria, we used data from 470 (47.9% men, 49.5% women, 2.4% other) of 512 participants in our analyses with a mean age of 31.2 ($SD = 11.4$). The education level ranged from graduate work (17.7%), bachelor's degree (32.6%), associate degree (8.5%), some college (25.9%),

[13] https://prolific.co/

[14] https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower

Figure 8

Political orientation of survey participants ranging from very liberal to very conservative



vocational or technical school (12.9%), high school graduate (1.19%), some high school (1.0%), to 8th grade (0.2%). On average, participants indicated a political orientation score of 3.4 (SD = 2.4; 0 = very liberal, 10 = very conservative). Figure 8 shows the respective distribution indicating a sample with a slightly liberal leaning. The size of our survey groups was balanced, with 158 subjects assigned to the control group and 160 individuals in each of the intervention groups.

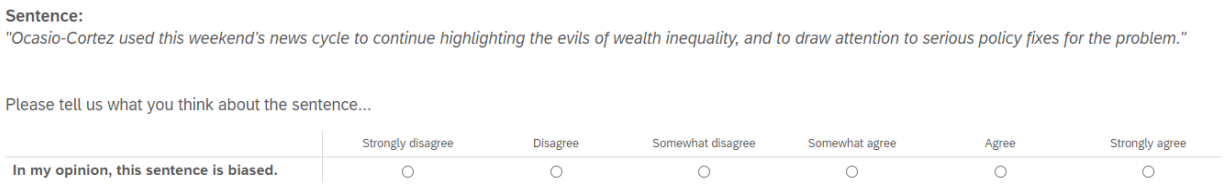
Measures

Perception of media bias in news sentences

Over both survey phases, we exposed participants to 46 sentences. We asked how biased the sentence was perceived by the subject measured on a 6-point Likert scale for every sentence. A sample question is shown in Figure 9. We explicitly did not provide a neutral option in the answer format to binarize the data. Initially, the 6-point Likert scale format was chosen to compute the inter-rater agreement within the groups for each sentence as an alternative explorative way to assess labeling accuracy without a gold standard. Due to time constraints, this analysis will be implemented in future research.

Figure 9

A sample question from the survey presenting a sentence and asking for the subject's media bias perception



Attention check

We included an attention check in both survey phases to ensure that participants paid attention and read the presented sentences carefully. The attention check item is shown in Figure 10. It was designed to resemble other survey items and not attract attention when clicking inattentively through the survey. The item was shown after completing the first half within the respective survey phase.

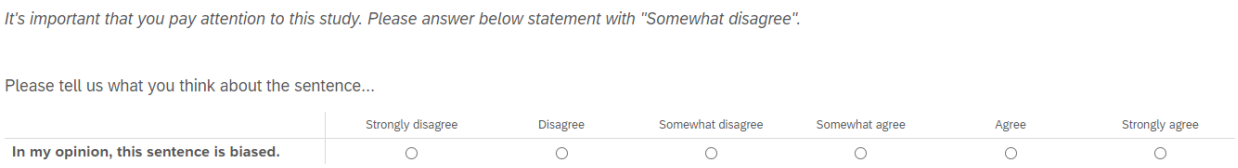
Accuracy of rating media bias

Our overall aim in terms of measurement comprises the calculation of a subject's media bias labeling accuracy on sentence level in all groups and survey phases to ascertain if human- and/or machine-generated bias annotations foster media bias awareness.

Since our data sample is not evenly distributed in terms of biased and non-biased sentences, calculating raw accuracy scores for the binary bias labeling task might lead to distorted results. Consider the following example pointing out in which cases classification accuracy is an unreliable performance metric: we let a participant annotate ten sentences with the labels *Biased* and *Non-biased* (Binary classification). The respective gold-standard labels exhibit an imbalanced class distribution, with eight sentences being biased and two non-biased sentences. If the participant now labels all sentences as biased, he/she achieves a reasonable accuracy of 80%, although he/she did not distinguish between biased and non-biased sentences. Thus, the raw accuracy score could lead to

Figure 10

The attention check item presented to participants after completing the first half of the train/test phase



misinterpretations regarding the participant's performance in detecting biased language.

An appropriate metric in case of imbalanced classification problems is the $F1$ score[15]. It can be considered the harmonic mean between *Precision* and *Recall* with a value range of $[0, 1]$, whereby larger scores indicate higher classification performance. Precision P is defined as the number of true positive classifications (T_p) proportionally to the number of false positive classifications (F_p): $P = \frac{T_p}{T_p+F_p}$. The Recall score R represents the ratio of true positive to false negative (F_n) classifications: $R = \frac{T_p}{T_p+F_n}$. $F1$ balances Precision and Recall by calculating their harmonic mean as follows:

$$F1 = 2 * \frac{P * R}{P + R} \tag{1}$$

Thus, $F1$ incorporates false positives and false negatives evenly and can be considered an appropriate performance metric for imbalanced classification problems (Brownlee, 2020).

We calculate the $F1$ score for every participant as follows: we split collected data according to the survey phases and survey groups. For every group and survey phase, we pre-process the data by binarizing the 6-point Likert scale bias items where the values *Strongly disagree*, *Disagree*, and *Slightly disagree* are encoded as *Non-biased* and *Strongly Agree*, *Agree*, and *Slightly agree* are categorized as *Biased*. Subsequently, we calculate a

[15] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html

subject's $F1$ score for the teaching and test phase based on the acquired binarized bias items compared to the gold-standard binary labels from Spinde, Plank, et al. (2021). In total, we thus get $n * m$ (n = number of participants, m = number of survey phases) $F1$ scores for every survey group used for our analyses of variance. In the remainder of the present work, we will use the term accuracy instead of the $F1$ score to keep descriptions as simple as possible.

Results

Statistical analyses

All computations that are part of the present thesis were performed with the open-source statistical package pingouin[16] written in the programming language Python. We started with analyzing descriptive statistics on the participants' labeling accuracy over all groups and survey phases. Then, we checked assumptions for calculating a two-way mixed ANOVA. We implemented the three labeling (human/machine/control) x two time point(teaching/test) mixed ANOVA to analyze the variance of the achieved labeling accuracy over all groups (between effect) and time points (within effect). Additionally, we computed pairwise t-tests (Sidak corrected) to check our experimental manipulations and hypotheses. The project and all analyses are preregistered on https://osf.io/95ht6/?view_only=19f054032f1d40b8ade5fcb42f78c568.

Descriptives on participants' accuracy in detecting sentence-level media bias

We measured a subject's accuracy of detecting media bias on sentence level by calculating it's $F1$ score over all sentences in both survey phases, respectively. The $F1$ score was our target metric for the subsequent analyses of variance showing if our bias visualizations foster media bias awareness. Table 1 represents means and standard deviations of participants' $F1$ scores in detecting sentence-level media bias, over all survey groups and phases. Figure 11 shows the respective distribution of $F1$ scores.

---

[16] https://pingouin-stats.org/#

Table 1

*Mean and standard deviations of participants' F*1 *scores over all survey groups and phases*

Group & Phase	*M*	*SD*
Control group - Train	0.619	0.13
Control group - Test	0.661	0.14
Intervention 1 - Train	0.709	0.14
Intervention 1 - Test	0.700	0.13
Intervention 2 - Train	0.648	0.13
Intervention 2 - Test	0.691	0.12

Assumptions for mixed ANOVA

We checked the following assumptions to calculate and interpret the mixed ANOVA.

Homogeneity of variance

The homoskedasticity check was performed based on a Levene Test testing the null hypothesis of equal variances. Table 2 shows the Levene test statistic W for both survey phases indicating equal variances.

Normality of residuals

We calculated the Shapiro-Wilk test statistic for all survey group-phase combinations to check if residuals are normally distributed. Table 3 shows the resulting test statistic.

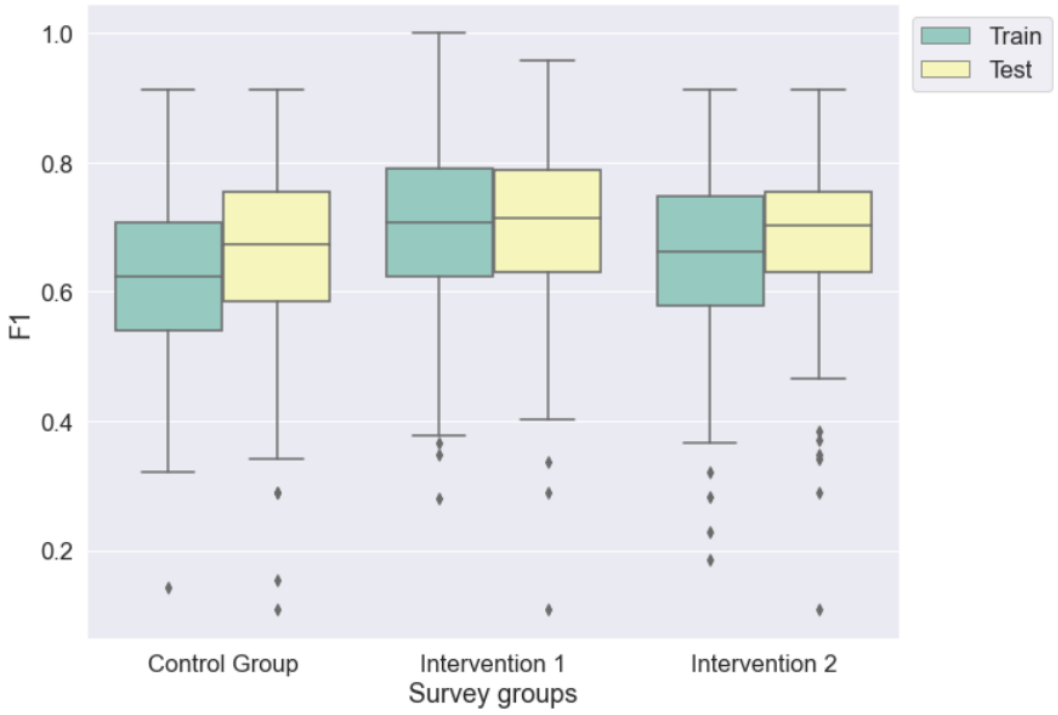
Homogeneity of covariances

The homogeneity of covariances was checked with a Box M test testing the null hypothesis of equal covariances. The test returned $\chi^2 = 1.61$ ($df = 2$, $p = 0.45$) indicating equal covariances.

Since the assumption of homoskedasticity and homogeneity of covariances were met, and only some group-phase combinations lacked of normally distributed residuals, we

Figure 11

Distribution of participants' $F1$ scores over all survey groups and phases



calculated and interpret a mixed ANOVA.

Effects of bias visualizations on detecting sentence-level media bias

We performed a two-way mixed ANOVA with three between factor and two within factor levels to examine general mean differences (based on $F1$ score) of detecting sentence-level media bias over all groups and phases. The between factor levels refer to the survey groups - a control group and two interventions groups. The within factor represents the train and test phase. Table 4 shows the resulting ANOVA output.

The $F1$ score means of detecting sentence-level media bias vary significantly over the survey groups ($F(2, 469) = 14.49$, $p = 7.81e^{-}7^{***}$, $\eta_p^2 = 0.058$) and survey phases ($F(2, 469) = 12.47$, $p = 4.54e^{-}4^{***}$, $\eta_p^2 = 0.026$). Thus, we observe a significant main effect of our experimental manipulations over the survey groups and survey phases. Furthermore, we observe a significant interaction of the between and within factor: $F(2, 469) = 5.73$, $p = 0.003^{**}$, $\eta_p^2 = 0.024$.

Table 2

Levene test statistic for both survey phases

Phase	W	p	Equal Variance
Train	0.29	0.75	True
Test	1.82	0.16	True

Table 3

Shapiro-Wilk test statistic for both survey phases

Group & Phase	W	p	Normally distributed
CG-Train	0.985795	0.104	True
CG-Test	0.941563	3.864e-06	False
I1-Train	0.981724	0.036	False
I1-Test	0.932176	8.617e-07	False
I2-Train	0.962731	3.099e-04	False
I2-Test	0.927323	3.896e-07	False

Simple effects analyses to test the experimental manipulation

We calculated pairwise post-hoc t-tests (Sidak corrected) to examine mean differences between single group pairs. Thereby, we could test our initial hypotheses. We performed one-sided tests for all directional hypotheses (H1 and H3), and for H2 (non-directional), we used two-sided tests. Table 5 shows test statistics for all pairwise t-tests.

Hypothesis 1

H1 states that both intervention groups achieve a significantly higher accuracy score ($F1$ score) in rating sentences in terms of bias than the control group in both phases. Intervention 1 (I1) and the control group (CG) as well as Intervention 2 (I2) and CG differ

Table 4

Output of two-way mixed ANOVA with three between (survey groups) and two within (survey phases) levels

Factors	SS	$DF1$	$DF2$	MS	F	p	η_p^2
Survey groups	0.65	2	469	0.32	14.49	$7.81e^{-}7$***	0.058
Phase	0.15	2	469	0.15	12.47	$4.54e^{-}4$***	0.026
Interaction	0.14	2	469	0.07	5.73	0.003**	0.024

Note. *$p < .05$, **$p < .01$, ***$p < 0.001$

significantly when ignoring the factor Phase (I1 - CG: $t = -4.99$, $p_{corr} = 1.51e^{-}6$***, $d = 0.56$; I2 - CG: $t = -2.54$, $p_{corr} = 0.017$*, $d = 0.29$). However, when distinguishing between survey group differences within the survey phases, we only observe a significant difference between CG and I1, containing bias visualizations based on human labels (Test phase: $t = -2.54$, $p_{corr} = 0.034$*, $d = 0.29$; Train phase: $t = -5.99$, $p_{corr} = 1.70e^{-}8$***, $d = 0.67$). I2 (machine-labeled bias visualizations) does not differ significantly from CG (Test phase: $t = -2.04$, $p_{corr} = 0.119$, $d = 0.23$; Train phase: $t = -1.96$, $p_{corr} = 0.142$, $d = 0.22$). Thus, only participants assigned to I2 achieved a significantly higher accuracy score in detecting media bias on sentence level than the control group, when distinguishing between the survey phases.

Hypothesis 2

H2 argues that both intervention groups do not differ significantly regarding their labeling accuracy in both the training and test phase. In the test phase, I1 and I2 do not differ significantly ($t = 0.59$, $p_{corr} = 0.993$, $d = -0.07$). However, when ignoring the information from the factor Phase, we observe significant differences between I1 and I2 ($t = 3.07$, $p_{corr} = 0.007$**, $d = -0.34$). Also, we find significant group differences in the train phase ($t = 4.09$, $p_{corr} = 3.29e^{-}4$***, $d = -0.46$).

Table 5

Pairwise t-tests for simple effects analyses between single group-phase pairs

Contrast	Phase	A	B	T	Alternative	p_{unc}	p_{corr}	d
Phase	-	Test	Train	3.50	one-sided	$9.99e^{-}1$	-	-0.19
Groups	-	CG	I1	-4.99	one-sided	$5.05e^{-}7$***	$1.51e^{-}6$***	0.56
Groups	-	CG	I2	-2.54	one-sided	0.006**	0.017*	0.29
Groups	-	I1	I2	3.07	two-sided	0.002**	0.007**	-0.34
Groups * Phase	Test	CG	I1	-2.54	one-sided	0.006**	0.034*	0.29
Groups * Phase	Test	CG	I2	-2.04	one-sided	0.021*	0.119	0.23
Groups * Phase	Test	I1	I2	0.59	two-sided	0.558	0.993	-0.07
Groups * Phase	Train	CG	I1	-5.99	one-sided	$2.84e^{-}9$***	$1.70e^{-}8$***	0.67
Groups * Phase	Train	CG	I2	-1.96	one-sided	0.025*	0.142	0.22
Groups * Phase	Train	I1	I2	4.09	two-sided	$5.49e^{-}5$***	$3.29e^{-}4$***	-0.46

Note. *$p < .05$, **$p < .01$, ***$p < 0.001$; d = Cohen's d

Hypothesis 3

H3 assumes an interaction effect between the bias visualizations and the survey phases on the participants' labeling accuracy. Table 4, showing the mixed ANOVA results, indicates a significant interaction term ($F = 5.73$, $p_{corr} = 0.003$**, $\eta_p^2 = 0.024$). Descriptives displayed in Table 1 as well as $F1$ score distributions shown in Figure 11 confirm that differences between intervention groups and the control group are smaller in the test phase.

Discussion

The present thesis investigated how human- and machine-generated bias labels and visualizations foster media bias awareness. Our results indicate that simple visual aids based on sentence-level bias highlighting increase a subject's performance of identifying slanted news coverage. However, this only accounts for visual aids based on human-generated bias labels. We believe that future scientific approaches dealing with

media bias benefit from our findings.

The first preparatory step for our media bias perception study comprised the development of BiasRoBERTa - a transformer-based Deep Learning model trained on biased and non-biased text (Pryzant et al., 2020). Then, we extracted a representative sample of 46 sentences from BABE (Spinde, Plank, et al., 2021) together with assigned binary bias labels. For our machine-based approach, we passed the unlabeled sentences to BiasRoBERTa and let the model assign bias labels. Our state-of-the-art model achieves 0.814 $F1$ score on BABE. Based on the human- and machine-generated bias labels, we highlighted biased sentences. Thereupon, we performed an online survey with two intervention groups (human- and machine-labeled sentences) and one control group to investigate the effects of the visualizations on the participants' media bias awareness. We divided the study into a train and test phase to investigate the visual aids' media bias teaching effect.

Our statistical analyses comprised descriptives on the participants' accuracy in detecting sentence-level media bias, a mixed ANOVA examining general group mean differences, and pairwise t-tests to check our hypotheses.

The descriptive analysis indicates $F1$ score variations over survey groups and phases with relatively stable standard deviations. In general, $F1$ scores tend to be higher in the intervention groups than in the control group. These tendencies can be observed in both the train and test phases. For example, the mean difference between I1/I2 and CG in the test phase is 3.9 and 3.0, respectively. The control group's means differ by 4.5 $F1$ points between the train and test phases. Although the difference seems to be quite substantial, we assume that subjects learned detecting media bias to some extent by reading and processing the sentences - a form of Experiential Learning (Kolb, 2014). However, additional Associative Learning promoted by our visual aids leads to overall higher $F1$ scores in the intervention groups. Future research could systematically test the Experiential Learning effect by plotting learning curves visualizing the participants'

performance course in detecting bias. An informative curve could be implemented based on moving average windows over both survey phases. A flattened curve would indicate that subjects do not substantially improve in the bias detection task.

Our two-way mixed ANOVA indicates significant main effects of our experimental manipulations and a significant interaction between the factors survey group and survey phase. Thus, we can observe substantial differences between the $F1$ means of our bias visualization groups and the control group as well as between the train and test phase. Due to the significant interaction term, we can also state that mean differences of the survey groups vary over survey phases. Results displayed in Table 1 and Figure 11 confirm that differences between intervention groups and the control group are smaller in the test phase. Accordingly, we can confirm Hypothesis 3, assuming an interaction effect between the survey groups and survey phases. The effect sizes of $\eta_p^2 = 0.058$ (survey groups), $\eta_p^2 = 0.026$ (survey phase),and $\eta_p^2 = 0.024$ (survey groups $*$ survey phase) can be considered as small[17].

Based on the pairwise t-tests' results shown in Table 5, we can partially confirm our first hypothesis stating that both human- and machine-based bias visualizations foster media bias awareness. Our first intervention group (I1) and the control group (CG) as well as our second intervention group (I2) and CG differ significantly when ignoring information from the factor Phase. However, when distinguishing between survey group differences within the survey phases, we only observe a significant difference between CG and I1, containing bias visualizations based on human labels. I2 (machine-labeled bias visualizations) does not differ significantly from CG. Thus, only participants assigned to I1 achieved a significantly higher $F1$ score in detecting media bias on sentence level than the control group when distinguishing between the survey phases. For all pairwise intervention-control group comparisons, we observe small to medium effect sizes (range $d =$ 0.22 - 0.56).

In general, these findings go along with Spinde, Jeggle, et al. (2021) reporting a

[17] According to https://www.statology.org/partial-eta-squared/

significant positive effect of in-text bias annotations on media bias awareness. However, our approach differs in two aspects:

First, we examined media bias perception on sentence-level (in contrast to article level in Spinde, Jeggle, et al. (2021) which is the plane of examination in the majority of media bias studies. Thereby, it was possible to incorporate automized bias detection approaches into a systematic online survey on media bias perception.

Second, our visual aids highlighting bias were substantially simpler than Spinde's word-level text annotations, including explanations on the present form of bias. However, the simplification did not lead to drastic effect size reductions. Our human-based bias visualizations achieved significant effects of similar sizes showing that complex visual aids might not be necessarily required to remind users of biased news coverage. For example, we achieved $d = 0.29$ for the comparison between I1 and CG in the test phase, indicating a small effect. Spinde, Jeggle, et al. (2021) report $\eta_p^2 = 0.025$ for the effect of their in-text annotations on media bias awareness, also indicating a small effect. However, we have to point out the different examination planes (article vs. sentence level). Further comparative research on the effectiveness of different visualization approaches would be necessary to infer more general conclusions.

Since we extracted gold-standard human-generated bias labels from Spinde, Plank, et al. (2021) for our human-generated bias visualizations in I1, and compared the clickworker's accuracy in detecting bias in both I1 and I2 with this benchmark, it seems plausible that our visual aids foster media bias awareness more substantially in I1 than in the machine-based I2. Considering that our machine-based bias detection approach labels a certain part of the sentences incorrectly ($F1 = 0.814$), we can assume that crowdsourcers learn erroneous associations between sentences and visualizations in some cases, which might explain the decreased performance improvement in identifying bias in I2 compared to I1.

A possible approach, testing if the performance gap between I1 and I2 is related to

using a human benchmark as evaluation ground-truth, could be the measurement of alternative accuracy scores. Based on the *Intraclass Correlation Coefficient (ICC)* measuring the agreement between raters on single sentences, we could derive an alternative form of bias labeling accuracy. Here, the examination units would no longer be participants but sentences. Using ICC as an accuracy metric would imply the advantage that human-labeled gold-standard bias data is not required for evaluation. However, a substantially larger sentence sample would be required to achieve adequate experimental power.

Next, we want to elaborate briefly on the participants' media bias awareness and resulting implications for future research. The subjects' accuracy scores in detecting sentence-level media bias in the test phase range from 0.66 $F1$ (CG) to 0.709 $F1$ (I1), indicating that identifying fine-grained linguistic bias in the news is a big challenge for non-experts. Subtle bias-inducing cues might be challenging to identify for both humans and machines. The finding coincides with Recasens and Jurafsky (2013) reporting low performance values in detecting word-level bias for both crowdsourcers and machines. Thus, future approaches incorporating automized bias detection and visualization techniques into bias perception surveys should be aware that bias labels assigned by non-experts are to be interpreted carefully.

Regarding our machine-based bias labeling approach, we want to point out BiasRoBERTa's acceptable performance of detecting sentence-level media bias ($F1 = 0.814$). However, visual aids based on binary bias labels assigned by the model did not significantly increase participants' media bias awareness. We merely observed a small non-significant effect. Further advancements in Natural Language Processing, such as the development of larger models capturing semantic and syntax more accurately, could be used to improve our bias classifier and the accuracy of resulting bias visualizations.

An additional future goal in media bias research could be the incorporation of sentence-level bias visualizations into a browser plugin. Built upon automatic bias

classification methods, bias visualizations could be created on the fly, reaching millions of news consumers.

Conclusion

Due to the continuous rise of worldwide online access to news of different quality, research on slanted news coverage becomes increasingly important. In the present thesis, we performed a first systematic study on the effects of human- and machine-generated media bias labels and visualizations on media bias awareness. Incorporating automized bias detection methods into psychological research is essential since human-based labeling approaches are not scalable to large amounts of news articles. Our results showed that simple human-based bias visualizations on sentence-level foster media bias awareness significantly. However, we observed small positive but non-significant effects for our machine-based approach. Our simple but effective visualizations might represent an essential building block of future automized bias annotation tools in consumer-oriented end products. Future tasks comprise replication studies based on larger sentence sample sizes, the improvement of automized bias classification methods, and the incorporation of alternative metrics to measure media bias awareness.

References

Aggarwal, S., Sinha, T., Kukreti, Y., & Shikhar, S. (2020). Media bias detection and bias short term impact assessment. *Array*, *6*, 100025. https://doi.org/10.1016/j.array.2020.100025

Alonso, H., Delamaire, A., & Sagot, B. (2017). Annotating omission in statement pairs, 41–45. https://doi.org/10.18653/v1/W17-0805

Ardèvol-Abreu, A., & Gil de Zúñiga, H. (2017). Effects 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. https://doi.org/10.1177/1077699016654684

Baron, D. P. (2006). Persistent media bias. *Journal of Public Economics*, *90*(1-2), 1–36. https://doi.org/10.1016/j.jpubeco.2004.10.006

Beltagy, I., Lo, K., & Cohan, A. (2019). Scibert: A pretrained language model for scientific text. https://doi.org/10.18653/v1/D19-1371

Bernhardt, D., Krasa, S., & Polborn, M. (2006). Political polarization and the electoral effects of media bias. *Journal of Public Economics*, *92*, 1092–1104. https://doi.org/10.1016/j.jpubeco.2008.01.006

Brownlee, J. (2020). *Imbalanced classification with python: Better metrics, balance skewed classes, cost-sensitive learning*. Machine Learning Mastery.

Budzianowski, P., & Vulić, I. (2019). Hello, it's gpt-2 – how can i help you? towards the use of pretrained language models for task-oriented dialogue systems. https://doi.org/10.18653/v1/D19-5602

Chen, W.-F., Al Khatib, K., Wachsmuth, H., & Stein, B. (2020). Analyzing political bias and unfairness in news articles at different levels of granularity. *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, 149–154. https://doi.org/10.18653/v1/2020.nlpcss-1.16

Cook, J., Lewandowsky, S., & Ecker, U. K. (2017). Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence. *PloS one*, *12*(5), e0175799. https://doi.org/10.1371/journal.pone.0175799

Cunningham, P., Cord, M., & Delany, S. J. (2008). Supervised learning. *Machine learning techniques for multimedia* (pp. 21–49). Springer. https://doi.org/10.1007/978-3-540-75171-7_2

D'Alessio, D., & Allen, M. (2000). Media bias in presidential elections: A meta-analysis. *Journal of Communication*, *50*, 133–156. https://doi.org/10.1111/J.1460-2466.2000.TB02866.X

Dallmann, A., Lemmerich, F., Zoller, D., & Hotho, A. (2015). Media bias in german online newspapers, 133–137. https://doi.org/10.1145/2700171.2791057

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. https://doi.org/10.18653/v1/N19-1423

Druckman, J. N., & Parkin, M. (2005a). The impact of media bias: How editorial slant affects voters. *The Journal of Politics*, *67*, 1030–1049. https://doi.org/10.1111/j.1468-2508.2005.00349.x

Druckman, J. N., & Parkin, M. (2005b). The impact of media bias: How editorial slant affects voters. *The Journal of Politics*, *67*(4), 1030–1049. https://doi.org/10.1111/j.1468-2508.2005.00349.x

Färber, M., Burkard, V., Jatowt, A., & Lim, S. (2020). A multidimensional dataset based on crowdsourcing for analyzing and detecting news bias. *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 3007–3014. https://doi.org/10.1145/3340531.3412876.

Festinger, L. (1957). *A theory of cognitive dissonance* (Vol. 2). Stanford university press.

Garrett, R. K. (2009). Echo chambers online?: Politically motivated selective exposure among internet news users. *Journal of computer-mediated communication*, *14*(2), 265–285. https://doi.org/https://doi.org/10.1111/j.1083-6101.2009.01440.x

Gururangan, S., Marasovic, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. *CoRR*, *abs/2004.10964*. https://doi.org/10.18653/v1/2020.acl-main.740

Hamborg, F., Donnay, K., & Gipp, B. (2018). Automated identification of media bias in news articles: An interdisciplinary literature review. *International Journal on Digital Libraries*, 1–25. https://doi.org/10.1007/s00799-018-0261-y

Han, X., & Eisenstein, J. (2019). Unsupervised domain adaptation of contextualized embeddings for sequence labeling. https://doi.org/10.18653/v1/D19-1433

Hastie, T., Tibshirani, R., & Friedman, J. (2009). Unsupervised learning. *The elements of statistical learning* (pp. 485–585). Springer. https://doi.org/10.1007/978-0-387-21606-5_14

Hochreiter, S. (1998). The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *6*, 107–116. https://doi.org/10.1142/S0218488598000094

Houston, J. B., Hansen, G., & Nisbett, G. (2011). Influence of user comments on perceptions of media bias and third-person effect in online news. *Electronic News*, *5*, 79–92. https://doi.org/10.1177/1931243111407618

Hube, C., & Fetahu, B. (2018). Detecting biased statements in wikipedia. *Companion proceedings of the the web conference 2018*, 1779–1786. https://doi.org/10.1145/3184558.3191640.

Kaye, B., & Johnson, T. (2016). Across the great divide: How partisanship and perceptions of media bias influence changes in time spent with media. *Journal of Broadcasting Electronic Media*, *60*, 604–623. https://doi.org/10.1080/08838151.2016.1234477

Klapper, J. T. (1960). The effects of mass communication.

Kolb, D. A. (2014). *Experiential learning: Experience as the source of learning and development*. FT press.

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology*. Sage publications.

Langendoen, D. (1964). Review of studies in linguistic analysis, ed. by j. r. firth. *Language*, *40*, 305–321. https://doi.org/10.2307/411592

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). Biobert: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234–1240. https://doi.org/10.1093/bioinformatics/btz682

Liddy, E. D. (2001). Natural language processing.

Lim, S., Jatowt, A., Färber, M., & Yoshikawa, M. (2020). Annotating and analyzing biased sentences in news articles using crowdsourcing. *Proceedings of the 12th Language Resources and Evaluation Conference*, 1478–1484. https://aclanthology.org/2020.lrec-1.184

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach.

Mitchell, A. (2014). *Political polarization & media habits*. Pew Research Center.

Mullainathan, S., & Shleifer, A. (2002). *Media bias* (Harvard Institute of Economic Research Working Papers No. 1981). Harvard - Institute of Economic Research. https://EconPapers.repec.org/RePEc:fth:harver:1981

Munson, S., & Resnick, P. (2013). Encouraging reading of diverse political viewpoints with a browser widget. *Proceedings of the 7th International Conference on Weblogs and Social Media, ICWSM 2013*.

Park, S., Kang, S., Chung, S., & Song, J. (2009). Newscube: Delivering multiple aspects of news to mitigate media bias. *Proceedings of CHI*, 443–452. https://doi.org/10.1145/1518701.1518772

Pavlov, I. P. (1949). Conditioned responses.

Pryzant, R., Martinez, R. D., Dass, N., Kurohashi, S., Jurafsky, D., & Yang, D. (2020). Automatically neutralizing subjective bias in text. *Proceedings of the aaai conference on artificial intelligence*, *34*(01), 480–489. https://doi.org/10.1609/aaai.v34i01.5385

Recasens, M., & Jurafsky, D. (2013). Linguistic models for analyzing and detecting biased language. *ACL 2013 - 51st Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, *1*.

Rojas, H. (2010). "corrective" actions in the public sphere: How perceptions of media and media effects shape political behaviors. *International Journal of Public Opinion Research*, *22*. https://doi.org/10.1093/ijpor/edq018

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020). Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter.

Schuldt, J. P., Konrath, S. H., & Schwarz, N. (2011). "global warming" or "climate change"? whether the planet is warming depends on question wording. *Public opinion quarterly*, *75*(1), 115–124. https://doi.org/10.1093/poq/nfq073

Siegel, S. (1983). Classical conditioning, drug tolerance, and drug dependence. *Research advances in alcohol and drug problems* (pp. 207–246). Springer.

Sindermann, C., Elhai, J. D., Moshagen, M., & Montag, C. (2020). Age, gender, personality, ideological attitudes and individual differences in a person's news spectrum: How many and who might be prone to "filter bubbles" and "echo chambers" online? *Heliyon*, *6*(1), e03214. https://doi.org/https://doi.org/10.1016/j.heliyon.2020.e03214

Spinde, T., Jeggle, C., Haupt, M., Gaissmaier, W., & Giese, H. (2021). How do we communicate media bias effectively? effects of bias visualizations.

Spinde, T., Krieger, J.-D., Ruas, T., Mitrović, J., Götz-Hahn, F., Aizawa, A., & Gipp, B. (2022). Exploiting transformer-based multitask learning for the detection of media bias in news articles. *Proceedings of the iConference 2022*. Retrieved March 4, 2022,

from https://media-bias-research.org/wp-content/uploads/2021/11/Spinde2022a_mbg.pdf

Spinde, T., Plank, M., Krieger, J.-D., Ruas, T., Gipp, B., & Aizawa, A. (2021). Neural media bias detection using distant supervision with babe - bias annotations by experts. *Findings of the Association for Computational Linguistics: EMNLP 2021*. https://doi.org/10.18653/v1/2021.findings-emnlp.101

Spinde, T., Rudnitckaia, L., Kanishka, S., Hamborg, F., Gipp, B., & Donnay, K. (2021). Mbic – a media bias annotation dataset including annotator characteristics. *Proceedings of the iConference 2021*. https://doi.org/10.6084/m9.figshare.17192924

Spinde, T., Rudnitckaia, L., Mitrovic, J., Hamborg, F., Granitzer, M., Gipp, B., & Donnay, K. (2021). Automated identification of bias inducing words in news articles using linguistic and context-oriented features. *Inf. Process. Manag.*, *58*, 102505. https://doi.org/10.1016/j.ipm.2021.102505

Spohr, D. (2017). Fake news and ideological polarization. *Business Information Review*, *34*, 150–160. https://doi.org/https://doi.org/10.1177/0266382117722446

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to fine-tune BERT for text classification? *CoRR*, *abs/1905.05583*. https://doi.org/10.1007/978-3-030-32381-3_16

Tsang, S. J. (2017). Cognitive discrepancy, dissonance, and selective exposure. *Media Psychology*, *22*, 1–24. https://doi.org/10.1080/15213269.2017.1282873

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 5998–6008.

Weeks, B. E., Lane, D. S., Kim, D. H., Lee, S. S., & Kwak, N. (2017). Incidental exposure, selective exposure, and political information sharing: Integrating online exposure patterns and expression on social media. *Journal of Computer-Mediated Communication*, *22*(6), 363–379. https://doi.org/10.1111/jcc4.12199

Williams, P., Kern, M. L., & Waters, L. (2016). Exploring selective exposure and confirmation bias as processes underlying employee work happiness: An intervention study. *Frontiers in Psychology*, *7*, 878. https://doi.org/10.3389/fpsyg.2016.00878

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., … Rush, A. M. (2020). Huggingface's transformers: State-of-the-art natural language processing.

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *Proceedings of the IEEE international conference on computer vision*, 19–27. https://doi.org/10.1109/iccv.2015.11