Neural Media Bias Detection Using Distant Supervision With BABE - Bias Annotations By Experts

Timo Spinde[1], Manuel Plank[2], Jan-David Krieger[2], Terry Ruas[1], Bela Gipp[1], Akiko Aizawa[3]

[1]University of Wuppertal, [2] University of Konstanz, [3] NII Tokyo
`{firstname.lastname}@{uni-wuppertal.de, @uni-konstanz.de}`
`aizawa@nii.ac.jp`

Abstract

Media coverage has a substantial effect on the public perception of events. Nevertheless, media outlets are often biased. One way to bias news articles is by altering the word choice. The automatic identification of bias by word choice is challenging, primarily due to the lack of a gold standard data set and high context dependencies. This paper presents BABE, a robust and diverse data set created by trained experts, for media bias research. We also analyze why expert labeling is essential within this domain. Our data set offers better annotation quality and higher inter-annotator agreement than existing work. It consists of 3,700 sentences balanced among topics and outlets, containing media bias labels on the word and sentence level. Based on our data, we also introduce a way to detect bias-inducing sentences in news articles automatically. Our best performing BERT-based model is pre-trained on a larger corpus consisting of distant labels. Fine-tuning and evaluating the model on our proposed supervised data set, we achieve a macro F_1-score of 0.804, outperforming existing methods.

1 Introduction

Online news articles have, over time, started to replace traditional print and radio media as a primary source of information (Dallmann et al., 2015). A varying word choice may have a major effect on the public and individual perception of societal issues, especially since regular news consumers are mostly not fully aware of the degree and scope of bias (Spinde et al., 2020a). As shown in existing research (Park et al., 2009; Baumer et al., 2015), detecting and highlighting media bias might be relevant for media analysis and to mitigate the effects of biased reports on readers. Also, the detection of media bias can assist journalists and publishers in their work (Spinde et al., 2021b). To date, only a few research projects focus on the detection and

aggregation of bias (Lim et al., 2020; Spinde et al., 2020c). Even though bias embodies a complex structure, contributions (Hube and Fetahu, 2019; Chen et al., 2020) often neglect annotator background and use crowdsourcing to collect annotations. Therefore, existing data sets exhibit low annotator agreement and inferior quality.

Our study holds both theoretical and practical significance. We propose BABE (**B**ias **A**nnotations **By E**xperts), a data set of media bias annotations, which is built on top of the MBIC data set (Spinde et al., 2021c). MBIC offers a balanced content selection, annotations on a word and sentence level, and is with 1,700 annotated sentences one of the largest data sets available in the domain. BABE improves MBIC, and other data sets, in two aspects. First, annotations are performed by trained experts and in a larger number. Second, the corpus size is expanded considerably with additional 2,000 sentences. The resulting labels are of higher quality and capture media bias better than labels gathered via crowdsourcing. In sum, BABE consists of 3,700 sentences with gold standard expert annotations on the word and sentence level.[1]

To analyze the ideal trade-off between the number of sentences, annotations, and human annotation cost, we divide our gold standard into 1,700 and 2,000 sentences, which are annotated by eight and five experts, respectively.[2] Lastly, we train and present a neural BERT-based classifier that outperforms existing approaches such as the one by Spinde et al. (2021b). Even though neural network architectures have been applied to the media bias domain (Hube and Fetahu, 2019; Chen et al., 2020), their data sets created using crowdsourcing do not exhibit similar quality as our expert data set. In addition, we include five state-of-the-art neural

---

[1]We also provide another 1,000 yet unlabeled sentences for future work. We have not labeled them to date due to resource restrictions.

[2]With the 1,700 stemming from MBIC (Spinde et al., 2021c).

models in our comparison and extend two of them in a distant supervision approach (Tang et al., 2014; Deriu et al., 2017). Leveraging large amounts of distantly labeled data, we formulate a pre-training task helping the model to learn bias-specific embeddings by considering bias information when optimizing its loss function. For the classification presented in this paper, we focus on sentence level bias detection, which is the current standard in related work (Section 2)[3]. We address future work on word level bias in Section 7. We publish all our code and resources on `https://github.com /Media-Bias-Analysis-Group/Neura l-Media-Bias-Detection-Using-Dis tant-Supervision-With-BABE`.

2 Related Work

Media bias can be defined as slanted news coverage or internal news article bias (Recasens et al., 2013). While there are multiple forms of bias, e.g., bias by personal perception or by the omission of information (Puglisi and Snyder, 2015), our focus is on bias caused by word choice, in which different words refer to the same concept. For a detailed explanation of the types of media bias, we refer to Spinde et al. (2021b). In the following, we summarize the existing literature on bias data sets and media bias classification.

2.1 Media Bias Data Sets

Lim et al. (2018) present 1,235 sentences labeled for word and sentence level bias by crowdsource workers. All the sentences in their data set focus on one event. Another data set focusing on just one event is presented by Färber et al. (2020). It consists of 2,057 sentences from 90 news articles, annotated with bias labels on article and sentence levels, and contains labels such as overall bias, hidden assumption, and framing. The annotators agree with a Krippendorff's α = -0.05. Lim et al. (2020) also provide a second data set with 966 sentences labeled on the sentence level. However, their reported interrater-agreement (IRR) of Fleiss' Kappa on different topics averages at zero.

Baumer et al. (2015) classify framing in political news. Using crowdsourcing, they label 74 news articles from eight US news outlets, collected from politics-specific RSS feeds on two separate days. Chen et al. (2020) create a data set of 6,964 arti-

cles containing political bias, unfairness, and non-objectivity labels at the article level. Altogether, they present 11 different topics such as "presidential election", "politics", and "white house".

Fan et al. (2019) present 300 news articles containing annotations for lexical and informational bias made by two experts. They define lexical bias as bias stemming from specific word choice, and informational bias as sentences conveying information tangential or speculative to sway readers' opinions towards entities (Fan et al., 2019). Their data set, BASIL, allows for analysis at the token level and relative to the target, but only 448 sentences are available for lexical bias.

Under the name MBIC, Spinde et al. (2021c) extract 1,700 sentences from 1,000 news articles. Crowdsource workers then label bias and opinion on a word and sentence level using a survey platform that also surveyed the annotators' backgrounds. MBIC covers 14 different topics and yields a Fleiss' Kappa score of 0.21.

Even though the referenced data sets contribute valuable resources to the media bias investigation, they still have significant drawbacks, such as (1) a small number of topics (Lim et al., 2018, 2020), (2) no annotations on the word level (Lim et al., 2018), (3) low inter-annotator agreement (Spinde et al., 2021c; Lim et al., 2020; Baumer et al., 2015; Lim et al., 2018), and (4) no background check for its participants (except (Spinde et al., 2021c)). Also, some related papers focus on framing rather than on bias (Baumer et al., 2015; Fan et al., 2019), and results are only partially transferable. Our work aims to address these weaknesses by gathering sentence level annotations about bias by word choice over a balanced and broad range of topics. The annotations are made by trained expert annotators with a higher capability of identifying bias than crowdsource workers.

2.2 Media Bias Classification Systems

Several studies tackle the automated detection of media bias (Hube and Fetahu, 2018; Spinde et al., 2020b; Chen et al., 2020). Most of them use manually created features to detect bias (Hube and Fetahu, 2018), and are based on traditional machine learning models (Spinde et al., 2021b).

Recasens et al. (2013) identify sentence level bias in Wikipedia using supervised classification. They use a bias lexicon and a set of various linguistic features (e.g., assertive verbs, sentiment)

---

[3]Our data set is in English language, which is also currently most common in the domain (Spinde et al., 2021a).

with a logistic regression classifier, identifying bias-inducing words in a sentence. They also report that crowdsource workers struggle to identify bias words that their classifier is able to detect.

Spinde et al. (2021b) create a media bias data set (i.e., MBIC) and develop a feature-based tool to detect bias-inducing words. The authors identify and evaluate a wide range of linguistic, lexical, and syntactic features serving as potential bias indicators. Their final classifier returns an F_1-score of 0.43 and 0.79 AUC. Spinde et al. point out the explanatory power of various feature-based approaches and the performance of their own model on the MBIC data set. Yet, their results indicate that Deep Learning models are promising alternatives for future work.

Hube and Fetahu (2018) propose a semi-automated approach to extract domain-related bias based on word embeddings properties. The authors combine bias words and linguistic features (e.g., report verbs, assertive verbs) in a random forest classifier to detect sentence level bias in Wikipedia. They achieve an F_1-score of 0.69 on a newly created ground truth based on Conservapedia.[4] In their following work, Hube and Fetahu (2019) propose a neural statement-level bias detection approach based on Wikipedia data. Using recurrent neural networks (RNNs) and different attention mechanisms, the authors achieve an F_1-score of 0.77, indicating a possible advantage of neural classifiers in the domain. Chen et al. (2020) train a RNN to classify article-level bias. They also conduct a reverse feature analysis and find that, at the word level, political bias correlates with categories such as negative emotion, anger, and affect.

To summarize, most approaches use manually created features, leading to lower performance and poor representation. The few existing contributions on neural models are based on naive data sets (cf. Section 2.1). Therefore, we decided to develop a neural classifier trained on BABE. Our system incorporates state-of-the-art models and improves their pre-training step through distant supervision (Tang et al., 2014; Deriu et al., 2017), allowing the model to learn bias-specific embeddings, thus improving its representation. Almost all models focus on sentence level bias, describing it as the lowest meaningful level that can be aggregated to higher levels, like the document level. Therefore, we follow the standard practice and construct a

sentence level classifier.

3 Data Set Creation

Since media bias by word choice rarely depends on context outside the sentences (Fan et al., 2019), we focused on gathering sentences only. To tackle the weaknesses of existing bias data sets, we created a robust and diverse corpus containing **B**ias **A**nnotations **B**y **E**xperts (BABE).

3.1 Data Collection

The general data collection and annotation pipeline is outlined in Figure 1. Similar to the filtering strategy proposed by Spinde et al. (2021b), the sentences should contain more biased than neutral sentences. BABE contains 3,700 sentences, 1,700 from MBIC (Spinde et al., 2021c) and additional 2,000. Like Spinde et al. (2021c), we extracted our sentences from news articles covering 12 pre-defined controversial topics.[5] The articles were published on 14 US news platforms from January 2017 until June 2020. We focused on the US media since their political scenario became increasingly polarizing over the last years (Atkins, 2016).
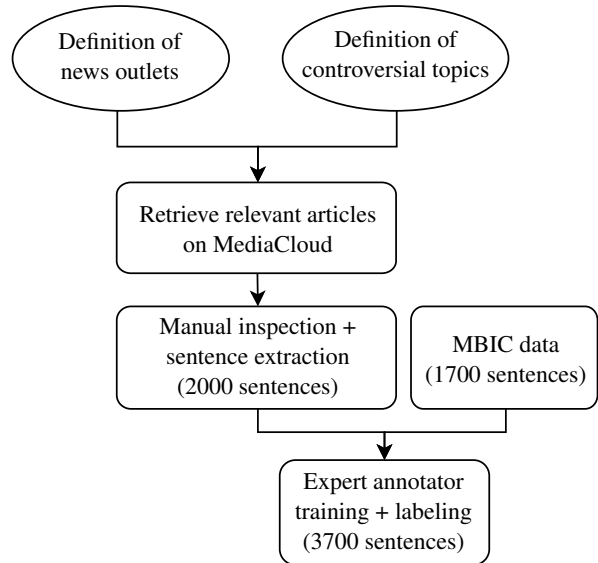

Figure 1: Data collection and annotation pipeline

We selected appropriate left-wing, center, and right-wing news outlets based on the media bias chart provided by Allsides.[6] The sentence collection was performed on the open-source media anal-

---

[4]https://conservapedia.com/Main_Page, accessed on 2021-04-10.

[5]The list of topics is provided at the repository mentioned in Section 1.

[6]https://www.allsides.com/media-bias/media-bias-chart, accessed on 2021-04-13.

ysis platform Media Cloud.[7] The collection process was as follows. We defined keywords describing every topic in one word or a short phrase, specified the news outlets, their time frame, and retrieved all available links for the relevant articles.[8] Then, we extracted sentences by manually inspecting the provided list of articles. The sentence selection was based on our media bias annotation guidelines comprising diverse examples of biased and neutral text instances (see Section 3.2).

3.2 Data Annotation

As laid out in Section 2, high-quality annotations are often obtained if the participants are properly instructed and have sufficient training (Fan et al., 2019; Spinde et al., 2021b). We compare our expert annotations with the crowdsourced labels provided by Spinde et al. (2021c) to further analyze quality differences between the two groups. Our results show that expert annotators render more qualitative bias labels than MBIC's crowdsourcers.

We define as an expert a person with at least six months of experience in the media bias domain and underwent sufficient training to (1) reliably identify biased wording, (2) distinguish between bias and plain polarizing language, and (3) take on a politically neutral viewpoint when annotating.[9] To build up such experience, we developed detailed instruction guidelines that are presented before the annotation task.[10] The instructions are substantially more comprehensive than instructions in a crowdsourcing setting. Considering that the annotation of bias on a fine-grained linguistic level is a complex task, and cognitive and language abilities likely have an impact on text perception (Kause et al., 2019), we hired only master students from programs completely held in English, who were among the top 20% with respect to their grade. Based on an iterative feedback loop between all annotators and us, we refined the guidelines multiple times with richer and clearer details. We discussed and evaluated existing annotations weekly as a group during the first three weeks of each annotator's work. We also always asked each annotator to hand in annotations before the discussion sessions, so they could not

influence each other. The annotators had to provide basic reasoning about their annotation decisions during our discussions. We maintained the labels only if the annotators were able to elaborate their annotations. Annotations of one annotator were discarded based on this method.

Apart from evaluation and instructions, each annotator rated at least 1,700 sentences to improve experience over time.[11] On average, per hour, they were paid 15,00€ and labeled 40 sentences, costing approximately 10,000€. The sum of money required to obtain a sufficient number of reasonable bias labels can be restrictive for media bias research. Therefore, BABE represents a major contribution that alleviates the lack of high-quality annotations in the domain. The annotators were instructed to label carefully and not as fast as possible, even though this resulted in a higher overall cost.

The general instructions for the annotation task were identical to the approach by Spinde et al. (2021c). First, raters were asked to mark words or phrases inducing bias. Then, we asked them to indicate whether the whole sentence was *biased* or *non-biased*. Lastly, the annotators labeled the sentence as *opinionated*, *factual, or mixed*.

As our resources were limited and the ideal trade-off between the number of sentences and annotators per sentence is not yet determined, we organized BABE into subgroups (SG), as described below:

- **SG1**. 1,700 sentences annotated by eight expert raters each.

- **SG2**. 3,700 sentences annotated by five expert raters each.

For SG1, we hired eight raters to annotate the 1,700 sentences (same as MBIC) on word and sentence levels (Spinde et al., 2021c).[12] Thereby, we obtained an expert-labeled ground truth comparable to MBIC's crowdsourcing results. For SG2, five of the previous eight annotators also labeled the 2,000 additionally collected sentences. We explored the ideal number of annotators by sampling. 5 annotators is a compromise between the agreement quality for both the bias and opinion labels, assuming that the annotation quality stays the same. To show the difference to 8 annotators, and as an outlook into future extensions of the data set, we

---

[7]https://mediacloud.org/, accessed on 2021-04-13.

[8]The keywords can be found at the repository mentioned in Section 1.

[9]Note: We cannot guarantee that a media bias expert is fully neutral, but we assume that an expert is able to leave political viewpoints aside to a substantial extent.

[10]Available on the repository mentioned in Section 1.

[11]The same sentences as in MBIC.

[12]In the original MBIC data set, each sentence was evaluated by ten crowdsource workers (Spinde et al., 2021c).

also release the codings made by 8 raters[13]. We will also add detailed statistics and results about all data and point out our selection process more clearly. As resources and time were limited, we leave the inclusion of further annotators and more sentences to future work. All raters were master students with a background in Data Science, Computer Science, Psychology, or Intercultural Communication. The groups and their annotators are described in detail in the repository mentioned in Section 1.

3.3 Evaluation of Data Sets

The raw labels obtained during the annotation phase were processed as follows. We calculated an aggregated bias/opinion label for every sentence based on a majority vote principle. For instance, if a sentence was labeled as biased by more than four expert annotators in SG1, we assigned the label *biased* to the sentence. Otherwise, the sentence was marked as *non-biased*.[14] The annotators did not agree on a label (no majority vote) in some sentences. Here, we assigned the label *no agreement*.

Our annotation scheme allows respondents to mark biased words. In SG1, a word is marked as biased if at least three annotators label it as such. In SG2, the threshold is subsequently reduced to two expert annotators labeling a word as biased.[15] We compute agreement metrics on the sentence level to acquire knowledge about data quality resulting from all annotation approaches. Our agreement metric choice is Krippendorff's α (Krippendorff, 2011), which is a robust agreement metric for studies including varying numbers of annotators per text instance (Antoine et al., 2014).

We first compared the annotations resulting from MBIC's crowdsourcing approach with our expert-based approach, including eight annotators labeling 1,700 sentences (SG1). Table 1 shows the agreement scores for the bias and opinion labels on a sentence level. Considering the bias agreement, SG1 exhibits fair agreement ($\alpha = 0.39$) and outperforms MBIC's agreement score ($\alpha = 0.21$).[16] A similar pattern can be observed regarding the opinion labels (i.e., SG1: $\alpha = 0.46$; MBIC: $\alpha = 0.26$). Furthermore, MBIC's crowdsourcers labeled more

Metric	Data	
	SG1	MBIC
Bias Agreement[1]	0.39	0.21
Opinion Agreement[1]	0.46	0.26
Total Biased Words	1530[3]	3283[3]
$\varnothing$ Biased Words [2]	1.95	2.40

[1] Calculated based on Krippendorff's α
[2] Average of bias words per biased sentence
[3] Out of 56,826 words in total

Table 1: Annotation results for the expert-annotated (SG1) and crowdsourced (MBIC) approach based on 1,700 sentences.

Label	Data	
	SG1	MBIC
Biased	43.88%	59.88%
Non-biased	47.05%	31.35%
No agreement	9.05%	8.76%
Opinionated	25.00%	30.65%
Factual	37.59%	33.65%
Mixed	26.64%	25.47%
No agreement	10.76%	10.24%

Table 2: Class distribution for SG1's and MBIC's 1700 sentences.

words as biased compared to SG1's experts, i.e., 3,283 vs. 1,530 (absolute) and 2.40 vs. 1.95 (average per biased sentence). Even though media bias detection is generally a difficult task, our inter-annotator agreement is much higher than in existing research in the domain, where α ranges between 0 and 0.20, as shown in Section 2.

Table 2 shows the label distribution comparison between SG1 and MBIC.[17] We can observe that our expert annotators (SG1) are more conservative in their annotation than the crowdsourcers (MBIC). In the expert data, 43.88% of the sentences are labeled as biased, whereas the crowdsources annotated 59.88%. The opinion labels' distribution is fairly balanced in both the expert annotator and crowdsourced data. Factual sentences occur slightly more often than opinionated sentences in both data sets.

Next, we evaluate our expert-based annotation approach, including five expert annotators labeling 3,700 sentences (SG2) in comparison to 1,700 (SG1). We compare metrics between both approaches to ascertain whether the reduced number of annotators in SG2 has a substantial impact on the annotator agreement. The finding could yield impli-

---

[13]But recommend to use 5-person ratings when using the full data set.

[14]Note: In SG2, the threshold reduced respectively due to the lower number of expert annotators.

[15]We manually inspected all instances to determine reasonable thresholds.

[16]The scoring interpretations are based on guidelines published by Landis and Koch (1977).

[17]Absolute numbers for all labels are reported in the code files at the repository mentioned in Section 1.

Metric	Data	
	SG1	SG2
Bias Agreement[1]	0.39	0.40
Opinion Agreement[1]	0.46	0.60

[1] Calculated based on Krippendorff's α

Table 3: Data set annotation results for the expert-based approaches (left: eight annotators labeling 1,700 sentences (SG1); right: five annotators labeling 3,700 sentences (SG2)).

Label	Data	
	SG1	SG2
Biased	43.88%	49.26%
Non-biased	47.05%	50.70%
No agreement	9.05%	0.00%
Opinionated	25.00%	23.35%
Factual	37.59%	43.54%
Mixed	26.64%	27.21%
No agreement	10.76%	5.88%

Table 4: Data set class distribution for the expert-based approaches (left: eight annotators labeling 1,700 sentences (SG1); right: five annotators labeling 3,700 sentences (SG2)).

cations for future research on our extended dataset (SG2). Table 3 shows agreement metrics for the bias and opinion labels of both expert-annotated approaches, and Table 4 represents label distributions. SG2 exhibits moderate agreement ($\alpha = 0.40$) in the bias annotation task, and slightly outperforms SG1 ($\alpha = 0.39$). Regarding the opinion labels, we observe a similar pattern, with SG2 outperforming SG1 more substantially (SG2: $\alpha = 0.60$; SG2: $\alpha = 0.46$). The expert annotators of SG1 are more conservative in labeling bias than SG2 (SG1: 43.88% vs. SG2: 49.26% labeled as biased).[18] The opinion labels are distributed marginally skewed in both annotator groups. Factual sentences occur more often than opinionated sentences in both data sets.

Further statistics on SG 1 and SG 2 such as bias/opinion distribution per news outlet and topic, the connection between bias and opinion, and the overall topic distribution are provided in the repository mentioned in Section 1.

---

[18]Due to the uneven number of annotators in SG2, "no agreement" cases do not exist here.

4 Methodology

We propose the use of neural classifiers with automated feature learning capabilities to solve the given media bias classification task. A distant supervision framework, similar to Tang et al. (2014), allows us to pre-train the feature extraction algorithms leading to improved language representations, thus, including information about a sample's bias. As obtaining large amounts of pre-training labeled data using humans is prohibitively expensive, we resort to noisy yet abundantly available labels that provide supervisory signals.

4.1 Learning Task

Given a corpus X and a randomly sampled sequence of tokens $x_i \in X$ with $i \in \{1, ..., N\}$, the learning task consists of assigning the correct label y_i to x_i where $y_i \in \{0, 1\}$ represents the *neutral* and *biased* classes, respectively. The supervised task can be optimized by minimizing the binary cross-entropy loss

$$\mathcal{L} := -\frac{1}{N} \sum_{i=1}^{N} \sum_{k=\{0,1\}} f_k(x_i) \cdot log(\hat{f}_k(x_i)). \quad (1)$$

where $f_k(\cdot)$ is a binary indicator triggering 0 in the case of neutral labels and 1 in the case of a biased sequence. $\hat{f}_k(\cdot)$ is a scalar representing the language model score for the given sequence.

4.2 Neural Models

We fit $\hat{f}_k(\cdot)$ using a range of state-of-the-art language models. Central to the architectural design of these models is Vaswani et al. (2017)'s encoder stack of the Transformer relying solely on the attention mechanism. Specifically, we use the BERT model (Devlin et al., 2019) and its variants DistilBERT (Sanh et al., 2019) and RoBERTa (Liu et al., 2019) that learned bidirectional language representations from the unlabeled text. DistilBERT is a compressed model of the original BERT, and RoBERTa uses a slightly different loss function with more training data than its predecessor. We also evaluate models built on the transformer architecture but differ in the training objective. While DistilBERT and RoBERTa use masked language modeling as a pre-training task, ELECTRA (Clark et al., 2020) uses a discriminative approach to learn language representations. We also include XLNet (Yang et al., 2019) in our comparison as an example of an autoregressive model. We systematically

evaluate the models' performance on the media bias sentence classification task. We also investigate the impact of an additional pre-training task introduced in the next section on the BERT and RoBERTa models' classification capabilities.

4.3 Distant Supervision

Fine-tuning general language models on the target task has proven beneficial for many tasks in NLP (Howard and Ruder, 2018). The language model pre-training followed by fine-tuning allows models to incorporate the idiosyncrasies of the target corpus. For text classification, the authors of ULMFiT (Howard and Ruder, 2018) demonstrated the superiority of task-specific word embeddings. Before fine-tuning, we introduce an additional pre-training task to improve feature learning capabilities considering media bias content. The typical unsupervised setting used in the general pre-training stage does not include information on language bias in the learning of the embedded space. To remedy this, we incorporate bias information directly in the loss function (equation 1) via distant supervision. In this approach, distant or *weak* labels are predicted from noisy sources, alleviating the need for data labeled by humans. Results by Severyn and Moschitti (2015) and Deriu et al. (2017) demonstrated that pre-training on larger distant datasets followed by fine-tuning on supervised data yields improved performance for sentiment classification.

A pre-training corpus is compiled consisting of news headlines of outlets with and without a partisan leaning to learn bias-specific word embeddings. The data source, namely, the news outlets, are leveraged to provide distant supervision to our system. As a result, the large amounts of data necessary to learn continuous word representations are gathered by mechanical means alleviating the burden of collecting expensive annotations. The assumption is that the distribution of biased words is denser in some news sources than in others. Text sampled from news outlets with a partisan leaning according to the Media Bias Chart is treated as biased. Text sampled from news organizations with high journalistic standards is treated as neutral. Thus, the mapping of bias and neutral labels to sequences is automated. The data collection resembles the collection of the ground-truth data described in Section 3. The defined keywords reflect contentious issues of the US society, as we assume slanted reporting to be more likely among those topics than

in the case of less controversial topics. The obtained corpus consisting of 83,143 neutral news headlines and 45,605 biased instances allows for the encoding of a sequence's bias information in the embedded space. The news headlines corpus serves to learn more effective language representations, it is not suitable for evaluation purposes due to its noisy nature. We ensure that no overlap exists between the distant corpus and BABE to guarantee model to guarantee model integrity with respect to training and testing.

5 Experiments

Training Protocol. We implement the neural models with HuggingFace's Transformer API (Wolf et al., 2020). The model components are instantiated with their pre-trained parameters. Parameters of the classification components are uniformly instantiated and learned. First, we fine-tune and evaluate neural models on BABE. Second, we identify the best performing model of the first run and include the distant supervision pre-training task.

Implementation. The hyperparameters remain unchanged for pre-training on the distant corpus and fine-tuning on BABE. Sentences are batched together with 64 sentences per mini-batch because estimating gradients in an online learning situation resulted in less stable estimates. To optimize $\mathcal{L}$, we use the Adam optimization with a learning rate of 5^{-5} (Kingma and Ba, 2014). Training on the distantly labeled corpus is performed for one epoch. While training on BABE, convergence can be observed after three to four epochs. A monitoring system is in place that stops training after two epochs without improvement of the loss and restores the parameters of the best epoch. All computations were performed on a single Tesla T4 GPU. All in all, pre-training and training of all models is executed in 5 hours.

Baseline. To assess the benefit of modern language models for the domain of media bias, we compare their performance to a traditional feature-based model (Baseline). We use the work by Spinde et al. (2021b) as our baseline method, as it offers the most complete set of features for the media bias domain. The authors use syntactic and lexical features related to bias words such as dictionaries of opinion words (Hu and Liu, 2004), hedges (Hyland, 2018) and assertive and factive verbs (Hooper, 1975). Spinde et al. (2021b)'s classifier serves as a baseline to evaluate our approach.

As feature-based models operate on the word level, we provide comparability by implementing the classification rule that the presence of a predicted biased word leads to the overall sentence being labeled as biased. In contrast, if the baseline model does not label words as biased in a given sequence, the sequence will be classified as neutral.

Evaluation Metric. Given the relatively small size of 3,700 sequences in BABE, we report performance metrics averaged on a 5 fold cross-validation procedure to stabilize the results. Because the class distribution in SG1 is slightly unbalanced, we use stratified cross-validation to preserve this imbalance in each fold. Following the standard in the literature, we report a weighted average of F_1-scores.

6 Results

Table 5 summarizes our performance results. Our baseline using engineered features exhibits low scores of 0.511 and 0.569 for SG1 and SG2, respectively.[19] BERT improves over the baseline by a large margin of 0.251 points on SG1 and 0.220 points on SG2. DistilBERT exhibits a lower performance for both corpora, whereas RoBERTa is the strongest representative of BERT-based models. Both models based on a different training approach than BERT, namely ELECTRA and XLNet, do not match the performance of BERT and its optimized variants. These results reaffirm established findings of the attention mechanism's advantage over traditional models (Hernández and Amigó, 2021) and indicate the benefits of large pre-trained models' for media bias detection.

Models trained and evaluated on SG2 generally perform better due to their bigger corpus size. The increase is around 0.02 points of the macro F_1-score for all models except RoBERTa + distant, where it is insignificant. Overall, we believe the improvement indicates that extending the data set in the future will be valuable.

Results of the fourth block of table 5 show that the distant supervision pre-training task leads to an improvement over BERT and RoBERTa. Our best performing model BERT + distant on SG2 achieves a macro F_1-score of 0.804 and improves over the BERT model by 0.02 points. Media bias can be better captured when word embedding algorithms are pre-trained on the news headlines corpus with distant supervision based on varying news outlets.

---

[19] In this Section, we show three decimal places to account for detailed model differences.

With the added data, information on a sequence's bias is incorporated in the loss function, which is not the case in "general purpose" language models.

Model	Macro F_1	
	SG1	SG2
Baseline	0.511 (0.008)	0.569 (0.008)
BERT	0.762 (0.019)	0.789 (0.011)
DistilBERT	0.758 (0.029)	0.777 (0.009)
RoBERTa	0.775 (0.023)	0.799 (0.011)
ELECTRA	0.742 (0.020)	0.760 (0.013)
XLNet	0.760 (0.042)	0.797 (0.015)
BERT + distant	0.778 (0.017)	**0.804** (0.014)
RoBERTa + distant	**0.798** (0.022)	0.799 (0.017)

Standard errors across folds in parentheses.
The first model block shows the best results of feature-based models. The second block of models consists of BERT and optimize variants. The models in the third block use new architectural or training approaches. The fourth block refers to models having learned bias-specific embeddings from the distantly supervised corpora.
The best results are printed in **bold**.

Table 5: Stratified 5 fold cross-validation results.

7 Discussion

Employing annotators with domain expertise allows us to achieve an inter-annotator agreement of $\alpha = 0.40$, which is higher than existing data sets (Spinde et al., 2021c). We believe domain knowledge and training alleviate the difficulty of identifying bias and are imperative to create a strong benchmark due to the complexity of the task. In future work, apart from improving the current data set and classifier, we will also explore why a text passage might be biased, not just its overall classification. Currently, traditional machine learning models are interpretable (Spinde et al., 2021b) but outperformed by recurrence and attention-based models. Hand-crafted features like static dictionaries cannot adequately address the complexity and context-dependence of bias.

We argue that standard metrics (e.g., accuracy and F_1) provide a limited perspective into a model's predictive power in case of a complex construct like media bias. Further research needs to tackle these pitfalls to propose systems with better generalization capabilities. A promising starting point might be a more refined evaluation scheme that decomposes the bias detection task into multiple sub-tasks, such as presented in CheckList (Ribeiro et al., 2020). This scheme will also allow us to understand how our system performs on different types of bias (e.g., bias by context, by linguistics, by overall reporting). Additionally,

we believe that current research on explainable artificial intelligence might increase users' trust in neural-based classifiers. Existing research already presents ways to visualize Transformer-based models and make their results more accessible and interpretable (Vig, 2019). Lastly, combining neural methods with advances in linguistic bias theory (Spinde et al., 2021b) to explain a classifier's decision to users will also be part of our future work.

For this work, we focused on sentence level bias, which is often used in the media bias domain. Still, in addition to the 3,700 labeled sentences, we also include word level annotations in our data set to encourage solutions focusing on more granular characteristics. We believe that word level bias conveys strong explanatory and structural knowledge and see a detailed word level bias analysis and detection as a promising research direction.

8 Conclusion

This work proposes BABE, a new high-quality media bias data set. BABE contains 3,700 labeled sentences, and enables us to compare crowdsourcing and expert annotations directly. Additionally, we propose a sentence level bias classifier based on BERT, which outperforms existing work in the domain. By deriving bias-specific word embeddings using distant supervision, we have improved our classifier even more, achieving a macro F_1-score = 0.804. We make all models, data, and code publicly available.[20]

Ethics/Broader Impact Statement

Detecting and highlighting media bias instances may have many positive implications and can mitigate the effects of such biases (Baumer et al., 2015). Still, bias is a highly sensitive topic, and some forms of bias especially rely on other factors than the content itself, such as a different perception of any text related to the individual background of a reader. When showing detected bias or news outlet classifications on a political or polarization scale to a reader, every algorithm should be transparent in how the classifications were made. In general, the topic should be handled carefully. We want to point out that it is uncertain if and how actual news consumers would like to obtain such information. Some research groups working on the detection of bias have also started to work on psychological and societal questions related to bias (Spinde et al.,

2020a). From a social science perspective, it remains to be explored how a classifier can mitigate the negative effects of biased media on society.

Generally, when performed in a balanced and transparent way, bias detection might positively affect collective decision-making and opinion formation processes. As such, and to this point, we see no immediate negative ethical or societal impacts of our work beyond what applies to other core building blocks of deep learning. Apart from the system transparency, as mentioned above, one important factor to consider when building, training, and presenting any media bias classifier is a manipulation protection strategy. Participants in any study, especially public ones, should not be able to tweak algorithms and therefore, e.g., flag neutral content as biased to undermine the validity of media bias detection systems. Hence, annotations should always be compared among multiple users, where trustworthiness can at least be largely assured. In open (crowdsourcing) scenarios, collecting user characteristics and consciously implementing specific content (like questions that should give an obvious answer but might be answered differently when users a following any pattern) is important.

As a side effect of our project, we experienced that our annotators learned to read the news more critically and reflected more about what they read even after the study ended. We have already started to implement the insights we gained into ways to improve the perception of bias in a game, teaching players to read news with greater care and execute a large study investigating how such a game can affect children, especially in school.

Our data set is completely anonymized to preserve the identities of everyone involved.

Acknowledgments

---

[20]We publish the link in Section 1.

References

Jean-Yves Antoine, Jeanne Villaneau, and Anaïs Lefeuvre. 2014. Weighted Krippendorff's alpha is a more reliable metrics for multi-coders ordinal annotations: experimental studies on emotion, opinion

and coreference annotation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 550–559, Gothenburg, Sweden. Association for Computational Linguistics.

Larry Atkins. 2016. *Skewed: A Critical Thinker's Guide to Media Bias*. G - Reference, Information and Interdisciplinary Subjects Series. Prometheus Books.

Eric Baumer, Elisha Elovic, Ying Qin, Francesca Polletta, and Geri Gay. 2015. Testing and comparing computational approaches for identifying the language of framing in political news. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1472–1482, Denver, Colorado. Association for Computational Linguistics.

Wei-Fan Chen, Khalid Al Khatib, Henning Wachsmuth, and Benno Stein. 2020. Analyzing Political Bias and Unfairness in News Articles at Different Levels of Granularity. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 149–154, Online. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR*.

Alexander Dallmann, Florian Lemmerich, Daniel Zoller, and Andreas Hotho. 2015. Media bias in german online newspapers. In *Proceedings of the 26th ACM Conference on Hypertext amp; Social Media*, HT '15, page 133–137, New York, NY, USA. Association for Computing Machinery.

Jan Deriu, Aurelien Lucchi, Valeria De Luca, Aliaksei Severyn, Simon Müller, Mark Cieliebak, Thomas Hofmann, and Martin Jaggi. 2017. Leveraging large amounts of weakly supervised data for multilanguage sentiment classification.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Lisa Fan, Marshall White, Eva Sharma, Ruisi Su, Prafulla Kumar Choubey, Ruihong Huang, and Lu Wang. 2019. In plain sight: Media bias through the lens of factual reporting. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6343–6349, Hong Kong, China. Association for Computational Linguistics.

Michael Färber, Victoria Burkard, Adam Jatowt, and Sora Lim. 2020. A multidimensional dataset based on crowdsourcing for analyzing and detecting news bias. In *Proceedings of the 29th ACM International Conference on Information Knowledge Management*, CIKM '20, page 3007–3014, New York, NY, USA. Association for Computing Machinery.

Adrián Hernández and José M. Amigó. 2021. Attention mechanisms and their applications to complex systems. *Entropy*, 23(3).

Joan B. Hooper. 1975. On assertive predicates. pages 91 – 124. Brill, Leiden, The Netherlands.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, page 168–177, New York, NY, USA. Association for Computing Machinery.

Christoph Hube and Besnik Fetahu. 2018. Detecting biased statements in wikipedia. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 1779–1786, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Christoph Hube and Besnik Fetahu. 2019. Neural based statement classification for biased language. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, WSDM '19, page 195–203, New York, NY, USA. Association for Computing Machinery.

Ken Hyland. 2018. *Metadiscourse: Exploring Interaction in Writing*, 1st edition. Bloomsbury Classics in Linguistics. Bloomsbury Academic, London.

Astrid Kause, Tarlise Townsend, and Wolfgang Gaissmaier. 2019. Framing climate uncertainty : Frame choices reveal and influence climate change beliefs. *Weather, Climate, and Society*, 11(1):199–215.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *arXiv e-prints*, page arXiv:1412.6980.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability. *Departmental Papers (ASC); University of Pennsylvania*.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Sora Lim, Adam Jatowt, Michael Färber, and Masatoshi Yoshikawa. 2020. Annotating and analyzing biased sentences in news articles using crowdsourcing. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1478–1484, Marseille, France. European Language Resources Association.

Sora Lim, Adam Jatowt, and Masatoshi Yoshikawa. 2018. Understanding characteristics of biased sentences in news articles. In *CIKM Workshops*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Souneil Park, Seungwoo Kang, Sangyoung Chung, and Junehwa Song. 2009. Newscube: delivering multiple aspects of news to mitigate media bias. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 443–452.

Riccardo Puglisi and James M. Snyder. 2015. Empirical studies of media bias. In Simon P. Anderson, Joel Waldfogel, and David Strömberg, editors, *Handbook of Media Economics*, volume 1 of *Handbook of Media Economics*, pages 647–667. North-Holland.

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

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of nlp models with checklist. In *Association for Computational Linguistics (ACL)*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Aliaksei Severyn and Alessandro Moschitti. 2015. UNITN: Training deep convolutional neural network for Twitter sentiment classification. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 464–469, Denver, Colorado. Association for Computational Linguistics.

Timo Spinde, Felix Hamborg, Karsten Donnay, Angelica Becerra, and Bela Gipp. 2020a. Enabling news consumers to view and understand biased news coverage: A study on the perception and visualization of media bias. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, JCDL '20, page 389–392, Virtual Event, China. Association for Computing Machinery.

Timo Spinde, Felix Hamborg, and Bela Gipp. 2020b. An integrated approach to detect media bias in german news articles. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, JCDL '20, page 505–506, Virtual Event, China. Association for Computing Machinery.

Timo Spinde, Felix Hamborg, and Bela Gipp. 2020c. Media bias in german news articles: A combined approach. *ECML PKDD 2020 Workshops: Workshops of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2020): INRA 2020, Ghent, Belgium, September 14–18, 2020, Proceedings*, 1323:581–590.

Timo Spinde, Christina Kreuter, Wolfgang Gaissmaier, Felix Hamborg, Bela Gipp, and Helge Giese. 2021a. Do you think it's biased? how to ask for the perception of media bias. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL)*.

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

Timo Spinde, Lada Rudnitckaia, Kanishka Sinha, Felix Hamborg, Bela Gipp, and Karsten Donnay. 2021c. MBIC – A media bias annotation dataset including annotator characteristics. In *Proceedings of the iConference 2021*. iSchools.

Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. 2014. Learning sentiment-specific word embedding for Twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1555–1565, Baltimore, Maryland. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Jesse Vig. 2019. A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42, Florence, Italy. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on*

Empirical Methods in Natural Language Processing: System Demonstrations, pages 38–45, Online. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.