Sentence-level Media Bias Analysis Informed by Discourse Structures

Yuanyuan Lei[1] **Ruihong Huang**[1] **Lu Wang**[2] **Nick Beauchamp**[3]
[1]Computer Science and Engineering, Texas A&M University, College Station, TX
[2]Computer Science and Engineering, University of Michigan
[3]Political Science, Northeastern University
[1]{yuanyuan, huangrh}@tamu.edu
[2]wangluxy@umich.edu, [3]n.beauchamp@northeastern.edu

Abstract

As polarization continues to rise among both the public and the news media, increasing attention has been devoted to detecting media bias. Most recent work in the NLP community, however, identify bias at the level of individual articles. However, each article itself comprises multiple sentences, which vary in their ideological bias. In this paper, we aim to identify sentences within an article that can illuminate and explain the overall bias of the entire article. We show that understanding the discourse role of a sentence in telling a news story, as well as its relation with nearby sentences, can reveal the ideological leanings of an author even when the sentence itself appears merely neutral. In particular, we consider using a functional news discourse structure and PDTB discourse relations to inform bias sentence identification, and distill the auxiliary knowledge from the two types of discourse structure into our bias sentence identification system. Experimental results on benchmark datasets show that incorporating both the global functional discourse structure and local rhetorical discourse relations can effectively increase the recall of bias sentence identification by 8.27% - 8.62%, as well as increase the precision by 2.82% - 3.48%[1].

1 Introduction

News media play a vast role not only by providing information, but also by selecting, packaging, and organizing the information to shape public opinions (Mccombs and Reynolds, 2002, 2009). Multiple studies showed that the media outlets are becoming more partisan and polarized, with great potential to influence public's political stance (Gentzkow and Shapiro, 2010b,a), which presents the necessity to develop novel models to detect the media bias.

Most recent research work focus on detecting the media bias either at the level of media outlet (Baly et al., 2018), or at the level of individual articles (Kiesel et al., 2019; Baly et al., 2020a; Roy and Goldwasser, 2020). However, each article itself comprises multiple sentences, which vary in their ideological bias (Entman, 2006, 2007a). In this paper, we focus on sentence-level media bias analysis to identify bias sentences that, as interpreted by (Gentzkow and Shapiro, 2006; Entman, 2007b; Mullainathan and Shleifer, 2002), provide the supportive or background information to shift opinion in an ideological direction, though that may be done via selective inclusion or omission as well as overt ideological language. The identified bias sentences can illuminate and explain the overall bias of the entire article. This is a difficult task, however, considering that ideological bias tends to be implicit and subtle, and a bias sentence itself can appear merely neutral.

While sophisticated semantic reasoning may be needed to determine if a sentence induces bias, we observe that understanding the discourse role of a sentence in news story telling can inform bias sentence identification. Specifically, we observe that sentences describing the main news event are less likely to carry bias, in contrast, certain types of supportive contents are more likely to induce bias, such as sentences describing reactions of various parties toward the main event or main entities. This is related to the selective reporting problem in news production (Broockman and Kalla, 2022; Enke, 2020), where journalists select what to include from among many materials that are all relevant to the main event to some extent, and the selected content can reveal the organization or journalist's leaning and stance on the main event or main entities.

Table 1 shows an example document, where the main event is *the democratic Rep. Braley "mocks" Grassley as farmer*. The author first introduced the main event in sentence one (S1) and continued to describe a followup event in S2 and S3, *Braley*

[1]The code link: https://github.com/yuanyuanlei-nlp/bias_sentence_discourse_emnlp_2022

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 10040 - 10050
December 7-11, 2022

	sentence text	discourse role
title	Dem running for Senate in Iowa mocks Grassley as farmer who never went to law school	
S1	Iowa Democratic Rep. Bruce Braley is under fire after a video surfaced Tuesday of him mocking Sen. Chuck Grassley as a "farmer from Iowa who never went to law school".	Main Event
S2	Braley apologized for the remarks in a written statement after the video was released.	Current Context
S3	He said he "respects" Grassley and proclaimed his support for Iowa's farmers.	Current Context
S4	A spokesperson for Grassley fired back Tuesday, saying that Braley as a trial lawyer is not be eligible to speak out on a number of policy on agriculture, energy, or healthcare.	Distant Evaluation
S5	*"Sen. Grassley is one of only two working family farmers in the United States Senate, where he brings Iowa common sense to work for, anti-trust, transportation, environmental, energy, trade, health care, communications, national security, and tax policy that works for all of America," the spokesperson said.*	Distant Evaluation

Table 1: An example document with five sentences and their news discourse roles shown on the right column. The fifth sentence (S5) is a bias sentence highlighted in italics and red.

	M1	M2	C1	C2	D1	D2	D3	D4	Total
bias	112 (12.70)	0 (0)	39 (11.34)	181 (12.35)	52 (12.29)	24 (**17.27**)	724 (**19.59**)	83 (8.99)	1222 (15.32)
nonbias	770 (87.30)	1 (100)	305 (88.66)	1285 (87.65)	371 (87.71)	115 (82.73)	2972 (80.41)	840 (91.01)	6755 (84.68)
bias	40 (31.50)	0 (none)	45 (**54.22**)	49 (**38.58**)	15 (31.25)	3 (**60.00**)	112 (31.55)	20 (25.32)	290 (34.44)
nonbias	87 (68.50)	0 (none)	38 (45.78)	78 (61.42)	33 (68.75)	2(40.00)	243 (68.45)	59 (74.68)	552 (65.56)

Table 2: Number (ratio) of bias and nonbias sentences under each of the eight news discourse role types for the BASIL (first two rows) and BiasedSents (latter two rows) datasets. The rightmost column shows the overall Number (ratio) of bias and nonbias sentences in an entire dataset. The discourse role wise ratios of bias sentences that are higher than the overall ratio are bolded. M1: Main Event, M2: Consequence, C1: Previous Context, C2: Current Context, D1: Historical Event, D2: Anecdotal Event, D3: Evaluation, D4: Expectation

made an apology, which was immediately triggered by the main event. Arguably, the first three sentences form a relatively complete news, and there is no clear opinion projected to either entity yet. But, the author continued to describe the reaction of *a spokesperson for Grassley* and included two quotations from this person, an indirect quotation (S4) and a direct quotation (S5), that commented on the two main entities. The long direct quotation (S5) proclaims the importance of Grassley's background and does cast a positive impression on the entity *Grassley*, especially when understood with respect to the main event of *Grassley* being mocked as farmer. Presumably, there were reactions from other parties or individuals toward this main event that are relevant to report as well, the fact that the author selected to include this particular individual's quotations reveal his ideology leaning.

In particular, we choose to incorporate discourse roles predicted by our recent system for news discourse profiling[2] (Choubey and Huang, 2021). The news discourse profiling task distinguishes three types of contents in a news article, main contents, context-informing contents and additional supportive contents, and labels each sentence with one of eight subtypes reflecting common discourse roles of a sentence in telling a news story. Specifically, 1). main contents have two subtypes, Main event (M1) and Consequence (M2), and cover sentences that describe the main event and their immediate consequences which are often found inseparable from main events. 2). context-informing contents have two subtypes, Previous Event (C1) and Current Context (C2), and cover sentences that explain the context or cause of the main event, including recent events and general circumstances, and 3). additional supportive contents have four subtypes, describing past events that precede the main event in months and years (Historical Event (D1)) or unverifiable situations that are often fictional or personal accounts of incidents of an unknown person (Anecdotal Event (D2)), or opinionated contents including reactions from immediate participants, experts, known personalities as well as journalists or news sources (Evaluation (D3)), except speculations and projected consequences that are labeled as Expectation (D4). Numerical analysis on two datasets (Table 2) show that depending on the dataset, a bias sentence is more commonly tagged as context-informing content or Anecdotal Event (D2) and Evaluation (D3) subtypes of additional supportive content.

[2]The system link: https://github.com/prafulla77/Discoure_Profiling_RL_EMNLP21Findings

no.	sentence text
title	Clinton Report Earnings of $139 Million in Seven Years
S1	Hillary Clinton on Friday released her most recent eight years of tax return.
S2	A month ago, Mr. Bush released his own tax returns and said he paid a higher tax rate than Clinton.
S3	Mr. Bush said his average federal tax rate was 36 percent, contrasted with Clinton's 30 percent rate.
S4	*It was unclear where Mr. Bush got the figure.*
S5	Mrs. Clinton's tax return showed that her federal tax rate was 35.72.

Table 3: Another example document. The fourth sentence (S4) is a bias sentence highlighted in italics and red, and this sentence is in a comparison relation with its previous sentence (S3).

In addition to global discourse roles in news story telling, we observe that local discourse relations with nearby sentences, the causal relation and the comparison relation in particular, can inform bias sentence identification as well. Causal relation implication is commonly used in journalism to attribute responsibility (Temmann et al., 2021). Interestingly, we found that the strong contrast semantics indicated by a comparison discourse relation can influence readers' perceptions of the related events or entities. In the example of table 3, the fourth sentence S4 itself has a neutral sentiment connotation, but when interpreted with respect to its previous sentence S3 and the comparison relation between them, this sentence has a purpose to challenge the authenticity of its previous sentence and has the effect to sway readers' opinions toward the event and involved entities. Therefore, we also train contingency and comparison discourse relation predictors using the PDTB corpus (Prasad et al., 2008) and incorporate their predictions to inform bias sentence identification.

We design a knowledge distillation model (Hinton et al., 2015) to distill the auxiliary knowledge from the global functional discourse role predictor and local rhetorical discourse relation predictors into our bias sentence detection system. An extra distillation loss is designed for guiding the detection model to learn from both global and local discourse structure predictors, so that the system can learn to take both discourse structures into account for building sentence representations. Specifically, a framework for response-based multi-teacher knowledge distillation is implemented, in which the student model takes predicted probability from the teacher model as learning material and aims to mimic teacher's behavior, as well as digests and integrates knowledge from multiple teachers. Experiments on two benchmark datasets (Fan et al., 2019; Lim et al., 2020) show that the knowledge distilled from both the global functional structure teacher and the local rhetorical structure teacher can increase the bias sentence identification recall by 8.27% - 8.62%,, as well as the precision by 2.82% - 3.48%, on top of a strong baseline system.

2 Related Work

Article-level media bias detection has attracted lots of attention in the nlp community (Hamborg et al., 2018). (Sapiro-Gheiler, 2019) utilized a text-based method for measuring news ideology. (Iyyer et al., 2014) used recurrent neural network for political ideology detection. (Baly et al., 2019) designed a multi-task original regression framework for jointly predicting the trustworthiness and the ideology of news media. (Liu et al., 2022) pretrained a language model for the political domain to better understand news political stance. (Baly et al., 2020b) proposed to prevent the model from learning media source as a shortcut for predicting ideology through an adversarial model. We, however, focus on detecting the media bias at the sentence-level.

Sentence-level media bias has a relatively short history in research. (Fan et al., 2019) is the first work to annotate bias sentences in a news document, and they also built a baseline model for sentence-level bias detection. (Lim et al., 2020) is another work annotating bias sentences with document contexts considered. (Spinde et al., 2021b,a) collected thousands of sentences from news articles and annotated them independent from the document they are taken from. Considering that bias sentences in a news article can be merely neutral or factual, sentence-level bias detection remains a challenging task.

News discourse is a news genre-specific discourse structure proposed by (Choubey et al., 2020), in which they categorized each sentence in a news article into eight types of discourse roles revolving around the main event. (Choubey and Huang, 2021) improved news discourse structure profiling

through an actor-critic framework, in which the explicit subtopic structure is used as critics and a combination model of the REINFORCE algorithm (Williams, 1992) and imitation learning (Hussein et al., 2017) is designed for training actor, and the interaction between sentences and the document is modeled in a hierarchy structure. In this paper, we used this state-of-art news discourse structure model as our teacher model.

PDTB discourse relation is provided by (Prasad et al., 2006), annotating explicit and implicit discourse relations between adjacent sentences or clauses in news articles. The newer version PDTB 2.0 (Prasad et al., 2008) added annotations of implicit relations across the entire corpus, and annotated sense of relations into four main classes: comparison, contingency, temporal, and expansion. PDTB 3.0 (Prasad et al., 2019) annotated additional implicit intra-sentential relations. As shown in (Liang et al., 2020), the sense-distribution of intra-sentential relations differs from that of inter-sentential relations. Considering that sentence-level media bias detection takes sentence as discourse unit, we used PDTB 2.0 data to train the teachers that predict local rhetorical discourse relations.

Knowledge distillation (Hinton et al., 2015) is a technique used for compressing large deep models as well as retaining its performance. Response-based knowledge distillation (Kim et al., 2018; Ba and Caruana, 2014; Mirzadeh et al., 2020) uses the soft logits of a large deep model as the teacher knowledge, and trains with a distillation loss to make student logits match teacher logits. Multi-teacher knowledge distillation (You et al., 2017; Lan et al., 2018; Song and Chai, 2018) utilized knowledge from different types of teachers and guide the student to build the ability of knowledge integration. In this paper, a response-based multi-teacher knowledge distillation framework is designed to distill two types of discourse structures.

3 The Distillation Model

In this section, we will elaborate on the bias sentence detection model distilling two types of discourse in detail. The model takes a whole news article consisting of N sentences $(S_1, S_2, \ldots, S_N)$ as input, and outputs the predicted probability of each sentence containing bias $(P_1^{bias}, P_2^{bias}, \ldots, P_N^{bias})$.

A framework of response-based multi-teacher knowledge distillation is designed, shown in Figure 1. Response, which is the soft probability predicted from the teacher model, is used as the learning materials for the bias detection model. An extra distillation loss is designed to guide the model to mimic the discourse teachers' response, so that the sentence embedding can be updated with two types of discourse informed as auxiliary knowledge.

3.1 Bias detection layers

RoBERTa (Liu et al., 2019) is utilized as the fundamental language model. The initial sentence embedding is the hidden state at the sentence start token <s>. Then a Bi-LSTM (Hochreiter and Schmidhuber, 1997) layer with the hidden dimension 384 is applied to capture the context information and derive the sentence embedding $(E_1, E_2, \ldots, E_N)$, in which $E_i \in R^d, i = 1, 2, \ldots, N$ and $d = 768$.

Two fully connected layers activated by the ReLU function are built on the top of sentence embeddings, as the *bias detection layers* to predict the probability of each sentence containing bias:

$$\begin{aligned} P_i^{bias} &= (p_i^{bias}, 1 - p_i^{bias}) \\ &= \sigma(W_2(ReLU(W_1 E_i + b_1)) + b_2) \end{aligned} \tag{1}$$

where $i = 1, 2, \ldots, N$, σ is the softmax function, $W_1 \in R^{d \times d}, W_2 \in R^{d \times 2}$

The loss for learning whether each sentence contains bias or not is the classical cross entropy loss:

$$\begin{aligned} L_{bias} = -\Big(\sum_{i=1}^{N} & y_i \log(p_i^{bias}) \\ & + (1 - y_i) \log(1 - p_i^{bias}) \Big) \end{aligned} \tag{2}$$

where $(y_1, y_2, \ldots, y_N)$ demonstrate the true label for each sentence, $y_i \in \{0, 1\}, i = 1, 2, \ldots, N$, value 1 means biased and 0 means unbiased.

3.2 Global discourse role prediction layers

The teacher model for the global functional discourse T_{global} is the current state-of-art news discourse structure model (Choubey and Huang, 2021). The teacher T_{global} classified eight discourse roles with an actor-critic framework, in which the explicit subtopic structure is used as critics, and a combination model of REINFORCE algorithm (Williams, 1992) and imitation learning is designed for training the actor.

The global discourse teacher T_{global} predicts the probability of eight discourse roles for each sentence $S_i, i = 1, 2, \ldots, N$ in the input article as

$$Q_i^{global} = (q_i^{global_1}, q_i^{global_2}, \ldots, q_i^{global_8})$$

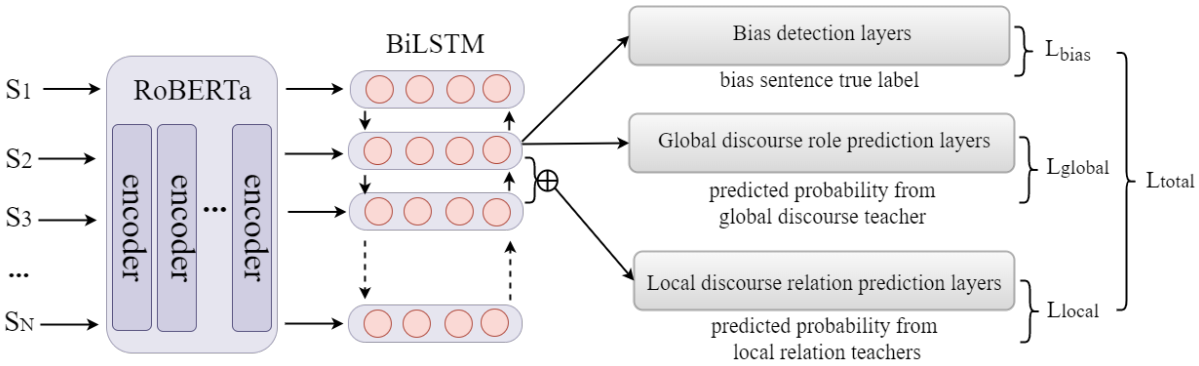

Figure 1: Bias detection model training informed by global and local discourse structures via knowledge distillation

Soft probability Q_i^{global} predicted by the teacher model is used as the learning material for global discourse structure.

Another two fully connected layers activated by the ReLU function are built on the top of sentence embeddings $E_i, i = 1, 2, \ldots, N$, as the student *global discourse role layers* for predicting the probability of eight discourse roles:

$$\begin{aligned} P_i^{global} &= (p_i^{global_1}, p_i^{global_2}, \ldots, p_i^{global_8}) \\ &= \sigma(W_4(ReLU(W_3 E_i + b_3)) + b_4) \end{aligned} \quad (3)$$

where $i = 1, 2, \ldots, N$, σ is the softmax function, $W_3 \in R^{d \times d}, W_4 \in R^{d \times 8}$. Soft probability P_i^{global} predicted by the student *global discourse role layers* represents its learning outcome.

The mean squared error loss between the predicted probability from the teacher Q_i^{global} and student layers P_i^{global} is designed to guide the student *global discourse role layers* to mimic the teacher's response, so as to learn the global discourse roles distilled from the teacher model T_{global}:

$$L_{global} = \sum_{i=1}^{N} \sum_{j=1}^{8} (p_i^{global_j} - q_i^{global_j})^2 \quad (4)$$

3.3 Local discourse relation prediction layers

The teacher models for the local *Comparison* and *Contingency* relations are both binary classification models and share the same model structure. Take the *Comparison* teacher as an example for illustration, the training data in PDTB dataset takes the sentence pair (Arg1, Arg2) as the input, and the label is 0 or 1 standing for whether the sentence pair has the comparison relation between them or not. RoBERTa (Liu et al., 2019) is the fundamental language model, and the concatenation of the hidden state at the sentence start token <s> of Arg1 and Arg2 is used as the feature vector. A fully connected layer is added on this feature vector to output the probability of whether comparison relation exist in (Arg1, Arg2) or not. Local *Comparison* and *Contingency* relation teacher are denoted as T_{comp} and T_{cont} respectively.

T_{comp}, T_{cont} predicts the probability of comparison / contingency for every adjacent sentence pairs $(S_{i-1}, S_i), i = 2, \ldots, N$ in the input article as

$$\begin{aligned} Q_i^{comp} &= (q_i^{comp}, 1 - q_i^{comp}) \\ Q_i^{cont} &= (q_i^{cont}, 1 - q_i^{cont}) \end{aligned} \quad (5)$$

Soft probability Q_i^{comp}, Q_i^{cont} predicted by the teacher model is used as the learning material for local discourse relations.

Fully connected layers activated by the ReLU function are added on the concatenation of the sentence embedding (E_{i-1}, E_i), as the student *local discourse relation layers* for predicting the probability of comparison / contingency:

$$\begin{aligned} P_i^{comp} &= (p_i^{comp}, 1 - p_i^{comp}) \\ &= \sigma(W_6(ReLU(W_5[E_{i-1}; E_i] + b_5)) + b_6) \\ P_i^{cont} &= (p_i^{cont}, 1 - p_i^{cont}) \\ &= \sigma(W_8(ReLU(W_7[E_{i-1}; E_i] + b_7)) + b_8) \end{aligned} \quad (6)$$

where $i = 2, \ldots, N$, σ is the softmax function, $W_5, W_7 \in R^{2d \times 2d}, W_6, W_8 \in R^{2d \times 2}$. Soft probability P_i^{comp}, P_i^{cont} outputted by the student *local discourse relation layers* represents its understanding of the two relations.

The cross entropy loss between the response Q_i^{comp}, Q_i^{cont} from the teacher, and the predicted probability P_i^{comp}, P_i^{cont} generated by the student layers, is additionally penalized to minimize the performance gap between the teacher model and student layers. In this way, local discourse relations

are distilled from the teacher T_{comp}, T_{cont} into the student *local discourse relation layers*:

$$\begin{aligned} L_{local} &= L_{comp} + L_{comp} \\ &= -\Big(\sum_{i=2}^{N} q_i^{comp} \log(p_i^{comp}) + (1 - q_i^{comp}) \log(1 - p_i^{comp})\Big) \\ &\quad -\Big(\sum_{i=2}^{N} q_i^{cont} \log(p_i^{cont}) + (1 - q_i^{cont}) \log(1 - p_i^{cont})\Big) \end{aligned} \tag{7}$$

3.4 The Learning Objective

The learning objective is the sum of bias detection layers loss, global discourse role layers loss, and the local discourse relation layers loss:

$$L_{total} = \lambda_{bias} L_{bias} + \lambda_{global} L_{global} + \lambda_{local} L_{local} \tag{8}$$

Learning the three types loss together can update the sentence embedding with two types of discourse incorporated. In this way, the global and local discourse structures are distilled as auxiliary knowledge into the bias sentence detection model.

4 Experiments

4.1 Datasets

The sentence-level bias detection task has a relatively short research history and few referable resources. BASIL and BiasedSents datasets are the only two available datasets till now that annotate bias sentences with context considered within a news article. Table 4 shows statistics of the two datasets.

BASIL dataset is the first work to annotate the sentence-level bias (Fan et al., 2019). It contains 100 triples of articles, each triple consists of three articles from three different media outlets discussing the same event, a total number of 300 articles. Fox News, New York Times, and Huffington Post are selected as the media outlets, and 10 sets are collected from each year between 2010 and 2019. The Cohen's kappa agreement between each annotator and the gold standard is from 0.34 to 0.70. The researcher demonstrates that bias sentences can be embedded uniformly across the entire article, and encoding contextual knowledge from the full articles is important.

BiasedSents dataset is another work of annotating news bias on sentence-level (Lim et al., 2020). It contains 46 articles from Sep 2017 to May 2018. They collected crowd-sourcing annotations in four scales: not biased, slightly biased, biased, and very biased. Following the same scenario of binary judgements (Lim et al., 2020), we also considered the first two scales as unbiased and the latter two as biased. The dataset provided the annotation from five different annotators, and we used the majority votes to derive the final gold labels. The Cohen's kappa agreement between each annotator and our gold label ranges from 0.17 to 0.58.

Dataset	# Articles	# Sentences	# Biased
BASIL	300	7977	1222
BiasedSents	46	842	290

Table 4: Statistics of BASIL and BiasedSents dataset

4.2 Teacher Models

We use the state-of-art model for news discourse profiling (Choubey and Huang, 2021) as out teacher model. We re-trained the model once using the same parameters described in the paper, and Table 5 shows its performance on the eight news discourse roles.

We trained our own teacher models for predicting contingency and comparison relations between sentences. The teacher models are both binary classification models and share the same simple architecture consisting of a fully connected layer applied on the concatenation of two sentence embeddings corresponding to two adjacent sentences. Followed the official suggestion in PDTB 2.0 dataset(Prasad et al., 2008), sections 2-21, sections 22 & 24 and section 23 are used for training, development and testing respectively. Both explicit and implicit relation data are utilized for training, because bias sentence may be in a discourse relation with neighbor sentences with or without a connectives explicitly shown. Table 6 shows the performance of *Comparison* and *Contingency* discourse relation teachers respectively.

4.3 Baseline Models

The previous work by (Fan et al., 2019) built a BERT (Devlin et al., 2019) baseline model for the bias sentence detection task. However, their model takes a single sentence as input and ignore the document context. In contrast, our distillation model takes the entire news article as the input, and make a prediction for each sentence in the input article.

Therefore, for fair comparison, in addition to a baseline model imitating the model in (Fan et al., 2019), we also built another baseline model that takes the entire news article consisting of N sen-

	M1	M2	C1	C2	D1	D2	D3	D4	Macro
Precision	56.30	33.33	28.69	57.63	66.76	52.98	65.45	57.61	57.10
Recall	49.57	24.68	25.35	58.83	60.34	51.15	68.77	65.19	55.34
F1-score	52.72	28.36	26.92	58.22	63.39	52.05	67.07	61.17	56.21

Table 5: Performance of global functional discourse teacher on NewsDiscourse dataset. M1: Main Event, M2: Consequence, C1: Previous Context, C2: Current Context, D1: Historical Event, D2: Anecdotal Event, D3: Evaluation, D4: Expectation

Comparison Relation	Contingency Relation
90.50 / 73.80 / 81.30	69.60 / 74.00 / 71.74

Table 6: Performance of Comparison relation teacher and Contingency relation teacher on the PDTB dataset, Precision / Recall / F1-score.

tences $(S_1, S_2, \ldots, S_N)$ as the input. To be detailed, the two baseline models are:

- **RoBERTa**: The hidden state at the sentence start token <s> of each sentence S_i is used as its sentence embedding. Then two fully connected layers activated by the ReLU function and a softmax layer are added on the top of sentence embedding as the *bias detection layers* to output the predicted probability.

- **RoBERTa + context**: Before the *bias detection layers*, a Bi-LSTM layer with the hidden dimension 384 is added on the hidden state at the <s> token, in order to encode context information when deriving sentence embeddings. This baseline model equals to our distillation model without discourse structure distilled.

4.4 Feature Concatenation Models

In addition to the two baseline models above, we present a feature concatenation model to incorporate the discourse structures as additional feature on top of *RoBERTa + context* model. For each sentence, we create a global discourse feature vector with eight probabilities for eight discourse roles predicted by the news discourse teacher model. Similarly, a local discourse feature vector consists of the probabilities for comparison and contingency relation with its adjacent sentences predicted by the PDTB teacher model. The global and local discourse structures feature vectors are concatenated with the sentence embedding in the *RoBERTa + context* model before the *bias detection layers*. Therefore, the feature concatenation model also incorporates the global and local discourse structures as additional information, but in a more naive way compared to the distillation model.

4.5 Experimental Setting

Ten-fold cross validation is performed, in each time, a fold is used as the test set, eight folds are used as the training set while a remaining fold is used as the validation set to determine when to stop training. Instead of spitting data into ten folders based on individual sentences as in (Fan et al., 2019), we split data based on articles. In our setting, sentences from the same article can never appear in both a training fold and a test fold, preventing the leaking of knowledge. After collecting the prediction results for the ten test folds, Precision, Recall, and F1-score of the bias class are reported.

The value of three λ hyper-parameters are obtained via grid search, in the range of [0,3] with a step size of 0.5.The value of λ_{bias} is set to be 1, and λ_{global} equals to 1.5, λ_{local} equals to 0.5. The training epochs is 5 for each testing task. We used AdamW (Loshchilov and Hutter, 2019) as the optimizer. The learning rate is adaptively adjusted by a linear schedular. The weight decay is set to be 1e-2. The dimension of sentence embedding, as well as the dimension of intermediate fully connected layers are set to be $d = 768$. We used Nvidia GeForce RTX 3090 for training the model. The running time of ten-folder cross validation is around two hours for our full model, and one hour for the baseline models.

4.6 Experimental Results

The results of 10-folder cross validation on the two datasets BASIL and BiasedSents are shown in Table 7. The first section of the table shows results of two baselines. Compared to *RoBERTa*, *RoBERTa + context* yields a little better performance across all the metrics, this shows that bias sentence identification benefits from having access to the wider context, however, the small improvements suggest that simply incorporating raw contexts with no focus or further analysis has limited effects.

	BASIL			BiasedSents		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Baseline Model						
RoBERTa	40.10	40.43	40.26	37.17	76.90	50.11
RoBERTa + context	40.71	41.57	41.13	38.10	77.24	51.03
Feature Concatenation Models						
+ Global News Discourse Structure	42.65	42.96	42.80	39.43	75.86	51.89
+ Local Discourse Relations	40.12	44.52	42.20	37.89	80.34	51.49
+ Both Global and Local	42.06	43.54	42.78	38.84	78.62	52.00
Distillation Models						
+ Global News Discourse Structure	43.41	46.64	44.97	41.42	76.55	53.75
+ Local Discourse Relations	43.06	46.48	44.71	37.50	**85.86**	52.20
+ Both (The Full Model)	**43.53**	**49.84**	**46.47**	**41.58**	85.17	**55.88**

Table 7: Ten-folder cross validation on bias sentences detection. Precision, Recall, F1 of bias class are shown. Both feature concatenation models and distillation models are built on the top of RoBERTa + context baseline model.

The second section of Table 7 shows results of adding one type or both types of discourse structures as additional features on top of the stronger baseline model *RoBERTa + context*. We can see that adding the additional features of global and local discourse structures yields consistent improvements on both precision and recall, which demonstrates the usefulness of the global and local discourse structures information. However, the amount of performance gain was not so impressive, up to 1.35% on precision and up to 1.97% on recall, which presents the necessity for having more sophisticated models in order to better incorporate the discourse structures.

The third section of Table 7 shows results of incorporating one type or both types of discourse structure information to the model *RoBERTa + context* by knowledge distillation. We can see that distilling the global discourse roles of sentences (the row *+Global News Discourse Structure*) improves the precision by 2.70% to 3.32%, as well as the recall by up to 5.07%. The improvements on both precision and recall metrics indicate that incorporating news discourse structures resolves both false-positive and false-negative predictions on bias sentences. Meanwhile, distilling the local comparison and contingency discourse relations (the row *+Local Discourse Relations*) can effectively seek out additional bias sentence that are otherwise overlooked by the contextualized RoBERTa baseline system (*RoBERTa + context*), and be able to improve the recall noticeably by 4.91% - 8.62%.

By comparing the distillation model results in the third section with the feature concatenation model results in the second section, we can see that the distillation models can better incorporate the global and local discourse structures, yielding extra gains of 1.47% - 2.74% in precision and 6.3% - 6.55% in recall.

Finally, the last row of Table 7 shows that the full system, when incorporating both global discourse role and local discourse relation information with the distillation model, can accumulate the performance gains and yields the best performance on both datasets. Compared to the baseline *RoBERTa + context*, the full model improves the recall of bias sentence identification by 8.27% and 8.62% on on BASIL and BiasedSents respectively, as well as improve the precision by 2.82% and 3.48% on the two datasets. The F1-score is improved by 4.85% - 5.34%. The significance test (the student t-test with 95% confidence level) shows that the full model significantly outperforms the baseline models with the p-value less than 2e-6.

4.7 Analysis

4.7.1 Global Functional Discourse

Here, we present the performance change across the eight discourse roles in Table 8. Incorporating the global functional discourse can bring precision and recall improvement across almost all the eight discourse roles, and help alleviate both false-negative and false-positive problems. The most improvement on Recall exist in the discourse role *Evaluation (D4)*, *Previous Context (C1)*, and *Current Context (C2)*, which is consistent with the analysis in Table 2 showing that these discourse roles contain more bias than other types. The most

	M1	M2	C1	C2	D1	D2	D3	D4	Overall
Precision	+ 9.01	nan	+ 5.23	+ 9.63	+ 5.14	- 1.68	+ 0.47	+ 1.52	+ 2.83
Recall	+ 10.71	nan	+ 2.56	+ 3.87	+ 1.92	+ 4.17	+ 10.64	- 1.20	+ 8.27
Precision	+ 1.30	nan	+ 1.27	+ 4.86	- 0.72	+ 0	+ 5.69	+ 4.89	+ 3.49
Recall	+ 0	nan	+ 2.22	+ 2.04	+ 1.33	+ 0	+ 10.71	+ 25.00	+ 7.93

Table 8: Precision and Recall change across eight global functional discourse roles in BASIL (first two rows) and BiasedSents (latter two rows). M1: Main Event, M2: Consequence, C1: Previous Context, C2: Current Context, D1: Historical Event, D2: Anecdotal Event, D3: Evaluation, D4: Expectation

	Comparison	Contingency
Precision	+ 3.37	+ 4.62
Recall	+ 0.68	+ 4.00
Precision	+ 3.74	+ 8.88
Recall	+ 5.26	+ 5.88

Table 9: Precision and Recall change in the instances with Comparison / Contingency relation in BASIL (first two rows) and BiasedSents (latter two rows).

improvement on Precision exist in the discourse role *Main Event (M1)*, *Current Context (C2)*, and *Expectation (D4)*, which is also consistent with the analysis in Table 2 that *Main Event* and *Expectation* contains less bias.

4.7.2 Local Rhetorical Discourse

Here, we study the effect of local discourse relations, and present the precision and recall change within the instances having comparison / contingency relation with nearby sentences in Table 9. We can see that all the evaluation metrics have the gain. Distilling the *Comparison* relation can improve the Recall for the instances with comparison relation by up to 5.26%. Distilling the *Contingency* relation can improve the Recall for the instances with contingency relation by up to 5.88%.

5 Conclusions

We study bias sentence identification within a news article, a challenging and important task that can illuminate and explain the overall bias of the entire article. We advocate for a discourse structure informed approach and have identified a global functional discourse structure and local rhetorical discourse relations as useful information for addressing this task. We also designed a knowledge distillation method that incorporates discourse structures and effectively informs bias sentence identification. For future work, we are keen to understand and categorize major mechanisms or strategies used by different news agencies to inject ideological bias.

6 Limitations

One major limitation is that we only experimented on English datasets. While both the global news discourse structure and local discourse relations we identified useful for bias detection in English news articles may also be useful for analyzing news articles written in other languages (more future work is needed to verify this), we acknowledge that it will not be easy to obtain the discourse structure teacher models our approach requires because there may not be relevant annotated corpora existing for other languages. To the best of our knowledge, the news discourse corpus does not have any non-English version yet, researchers have started to create PDTB-style discourse annotations for other languages, but the languages considered are still limited to several resource rich languages such as Arabic, Hindi and Chinese.

7 Ethical Considerations

As our evaluation shows, the presented bias sentence detection system has not achieved a satisfactory level of performance and may make false bias predictions when deployed.

Acknowledgements

We would like to thank the anonymous reviewers for their valuable feedback and input. We gratefully acknowledge support from National Science Foundation (NSF) via the awards IIS-2127746 and IIS-1942918. Lu Wang is supported through NSF grant IIS-2127747.

References

Jimmy Ba and Rich Caruana. 2014. Do deep nets really need to be deep? In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. 2020a. We can detect your bias:

Predicting the political ideology of news articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4982–4991, Online. Association for Computational Linguistics.

Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018. Predicting factuality of reporting and bias of news media sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3528–3539, Brussels, Belgium. Association for Computational Linguistics.

Ramy Baly, Georgi Karadzhov, Abdelrhman Saleh, James Glass, and Preslav Nakov. 2019. Multi-task ordinal regression for jointly predicting the trustworthiness and the leading political ideology of news media. *arXiv preprint arXiv:1904.00542*.

Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. 2020b. We can detect your bias: Predicting the political ideology of news articles. *arXiv preprint arXiv:2010.05338*.

David Broockman and Joshua Kalla. 2022. The manifold effects of partisan media on viewers' beliefs and attitudes: A field experiment with fox news viewers. *OSF Preprints*, 1.

Prafulla Kumar Choubey and Ruihong Huang. 2021. Profiling news discourse structure using explicit subtopic structures guided critics. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1594–1605, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Prafulla Kumar Choubey, Aaron Lee, Ruihong Huang, and Lu Wang. 2020. Discourse as a function of event: Profiling discourse structure in news articles around the main event. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5374–5386, Online. Association for Computational Linguistics.

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.

Benjamin Enke. 2020. What you see is all there is. *The Quarterly Journal of Economics*, 135(3):1363–1398.

Robert Entman. 2007a. Framing bias: Media in the distribution of power. *Journal of Communication*, 57:163 – 173.

Robert M. Entman. 2006. Framing: Toward Clarification of a Fractured Paradigm. *Journal of Communication*, 43(4):51–58.

Robert M. Entman. 2007b. Framing Bias: Media in the Distribution of Power. *Journal of Communication*, 57(1):163–173.

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.

Matthew Gentzkow and Jesse M Shapiro. 2010a. Ideological segregation online and offline. Working Paper 15916, National Bureau of Economic Research.

Matthew Gentzkow and Jesse M. Shapiro. 2010b. What drives media slant? evidence from u.s. daily newspapers. *Econometrica*, 78(1):35–71.

Matthew Gentzkow and Jesse M. Shapiro. 2006. Media bias and reputation. *Journal of Political Economy*, 114(2):280–316.

Felix Hamborg, Karsten Donnay, and Bela Gipp. 2018. Automated identification of media bias in news articles: an interdisciplinary literature review. *International Journal on Digital Libraries*, pages 1–25.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *ArXiv*, abs/1503.02531.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Ahmed Hussein, Mohamed Medhat Gaber, Eyad Elyan, and Chrisina Jayne. 2017. Imitation learning: A survey of learning methods. *ACM Comput. Surv.*, 50(2).

Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. 2014. Political ideology detection using recursive neural networks. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1113–1122, Baltimore, Maryland. Association for Computational Linguistics.

Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. 2019. SemEval-2019 task 4: Hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 829–839, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Jangho Kim, Seonguk Park, and Nojun Kwak. 2018. Paraphrasing complex network: Network compression via factor transfer. *ArXiv*, abs/1802.04977.

Xu Lan, Xiatian Zhu, and Shaogang Gong. 2018. Knowledge distillation by on-the-fly native ensemble. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 7528–7538, Red Hook, NY, USA. Curran Associates Inc.

Li Liang, Zheng Zhao, and Bonnie Webber. 2020. Extending implicit discourse relation recognition to the PDTB-3. In *Proceedings of the First Workshop on Computational Approaches to Discourse*, pages 135–147, Online. Association for Computational Linguistics.

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

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. *ArXiv*, abs/1907.11692.

Yujian Liu, Xinliang Frederick Zhang, David Wegsman, Nick Beauchamp, and Lu Wang. 2022. Politics: Pretraining with same-story article comparison for ideology prediction and stance detection. *arXiv preprint arXiv:2205.00619*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *ICLR*.

Maxwell Mccombs and A. Reynolds. 2009. How the news shapes our civic agenda. *Media Effects: Advances in Theory and Research*, pages 1–16.

Maxwell Mccombs and Amy Reynolds. 2002. *News Influence on Our Pictures of the World*, pages 1–18.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. 2020. Improved knowledge distillation via teacher assistant. In *AAAI*.

Sendhil Mullainathan and Andrei Shleifer. 2002. Media Bias. NBER Working Papers 9295, National Bureau of Economic Research, Inc.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K. Joshi, and Bonnie Lynn Webber. 2008. The penn discourse treebank 2.0. In *LREC*.

Rashmi Prasad, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind Joshi, and Bonnie Webber. 2006. The penn discourse treebank 1.0 annotation manual.

Rashmi Prasad, Bonnie Webber, Alan Lee, and Aravind Joshi. 2019. Penn Discourse Treebank Version 3.0.

Shamik Roy and Dan Goldwasser. 2020. Weakly supervised learning of nuanced frames for analyzing polarization in news media. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7698–7716, Online. Association for Computational Linguistics.

Eitan Sapiro-Gheiler. 2019. Examining political trustworthiness through text-based measures of ideology. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

Guocong Song and Wei Chai. 2018. Collaborative learning for deep neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 1837–1846, Red Hook, NY, USA. Curran Associates Inc.

Timo Spinde, Manuel Plank, Jan-David Krieger, Terry Ruas, Bela Gipp, and Akiko Aizawa. 2021a. Neural media bias detection using distant supervision with BABE - bias annotations by experts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1166–1177, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Timo Spinde, Lada Rudnitckaia, Kanishka Sinha, Felix Hamborg, Bela Gipp, and Karsten Donnay. 2021b. Mbic – a media bias annotation dataset including annotator characteristics.

Linn Julia Temmann, Annemarie Wiedicke, Sophia Schaller, Sebastian Scherr, and Doreen Reifegerste. 2021. A systematic review of responsibility frames and their effects in the health context. *Journal of Health Communication*, 26(12):828–838.

Ronald J. Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.*, 8(3–4):229–256.

Shan You, Chang Xu, Chao Xu, and Dacheng Tao. 2017. Learning from multiple teacher networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, page 1285–1294, New York, NY, USA. Association for Computing Machinery.