Sentence-level Media Bias Analysis with Event Relation Graph

Yuanyuan Lei and Ruihong Huang
Department of Computer Science and Engineering
Texas A&M University, College Station, TX
{yuanyuan, huangrh}@tamu.edu

Abstract

Media outlets are becoming more partisan and polarized nowadays. In this paper, we identify media bias at the sentence level, and pinpoint bias sentences that intend to sway readers' opinions. As bias sentences are often expressed in a neutral and factual way, considering broader context outside a sentence can help reveal the bias. In particular, we observe that events in a bias sentence need to be understood in associations with other events in the document. Therefore, we propose to construct an event relation graph to explicitly reason about event-event relations for sentence-level bias identification. The designed event relation graph consists of events as nodes and four common types of event relations: coreference, temporal, causal, and subevent relations. Then, we incorporate event relation graph for bias sentences identification in two steps: an event-aware language model is built to inject the events and event relations knowledge into the basic language model via soft labels; further, a relation-aware graph attention network is designed to update sentence embedding with events and event relations information based on hard labels. Experiments on two benchmark datasets demonstrate that our approach with the aid of event relation graph improves both precision and recall of bias sentence identification [1].

1 Introduction

Media bias refers to the ideological leaning of the journalists within news reporting (Lichter, 2017; Gentzkow and Shapiro, 2006a). News media plays an important role not only in providing information, but also in shaping public opinions (Van Leuven and Slater, 1991; Van Dijk et al., 1995; Strömbäck, 2012; Lei and Cao, 2023). Media outlets are exhibiting growing partisanship and polarization, with a greater potential to influence public opinions and interfere democratic process (Prior, 2013; Wilson et al., 2020; Lei et al., 2022). Thus, developing sophisticated models to detect media bias is important and necessary.

While media bias detection has received growing research interests, the majority of prior work detect media bias at the article level (Baly et al., 2020; Kiesel et al., 2019). Nevertheless, each sentence within an article serves for different purposes in narrating a news story. It is important to pinpoint those bias sentences that aim to implant ideological bias and manipulate public opinions. Bias sentence, as interpreted by Gentzkow and Shapiro (2006b); Entman (2007); Mullainathan and Shleifer (2002), is the content providing supportive or background information to shift pubic opinions in an ideological direction, though that may be done via information selective inclusion or omission as well as overt ideological language. This paper aims to identify media bias at sentence level, and develop computational model to detect bias sentences.

Identifying sentence-level media bias still remains a challenging task (Vargas et al., 2023), especially considering its subtle and implicit nature. Take the example article in Figure 1 as an instance, while the first bias sentence S5 contains overt ideological language and may be easily identified, the second bias sentence S6 looks completely neutral and factual, and a model merely examining the sentences in isolation without considering broader context may not readily reveal such bias cases. But it becomes clear that S6 carries bias if we contrast S6 (*Trump always decided not to follow through presidential selections*) with what Trump said in the statement (*his willingness to still run if he was displeased with the candidates*). Interestingly, it appears that the author hinted on the relatedness between the historical events described in S6 and the *statement* event by explicitly stating a temporal relation between them. To further elaborate, Trump claimed a causal relation between events *drop plans*

[1]The code and data link: https://github.com/yuanyuanlei-nlp/sentence_level_media_bias_naacl_2024

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 5225–5238

Example Article	
S1	Donald J. Trump said in a **statement** on Tuesday that he **dropped** plans to moderate a presidential debate.
S2	Trump said that he **dropped** because the Republican Party asked him to **give up** the right to run as an independent candidate if he moderates the debate.
S3	"It is very important to me that the right Republican candidate be chosen to defeat the failed Obama administration," Mr. Trump said in his **statement**.
S4	"But if that Republican, in my opinion, is not the right candidate, I am not willing to **give up** my right to run as an independent candidate."
S5	Mr. Trump drew considerable **criticism** from the Republican establishment and attracted the **ridicule** from the public for his decision to drop and his willingness to still run if he was displeased with the candidates.
S6	Before this statement, Mr. Trump has **run** for president several times since the 1980s but has always decided not to follow through – he chose to **drop** the debate every time.

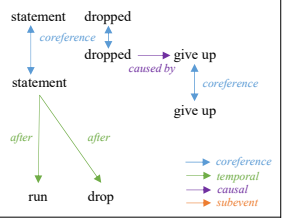


Figure 1: An example article containing bias sentences, and its corresponding event relation graph. Bias sentences are highlighted in red. Events words are shown in **bold** text. Event relation graph consists of events as nodes and four types of event relations: coreference, temporal, causal, and subevent relation.

to moderate a debate and *give up right as an independent candidate*, revealing the historical events in S6 propels the reader to rethink whether the stated cause is indeed the real reason, given his consistent past behavior of *dropping presidential election several times since 1980s*.

While it remains difficult to fully encode the intricate reasoning process used for recognizing sentence-level bias, inspired by the observed importance of understanding events in the context of other related events, we propose to construct an event relation graph and connect events in a document with four common types of event relations: coreference, temporal, causal and subevent relations. The event relation graph explicitly captures event-level content structures and enhances the comprehension of events within their interrelated context. We employ this event relation graph to guide the bias sentence detector and engage its attention on significant event-event relations.

Moreover, we propose to integrate the event relation graph into bias sentences identification in two ways. Firstly, an event-aware language model is trained using the soft labels derived from the event relation graph, thereby injecting the knowledge of events and event-event relations into the language model. Secondly, a relation-aware graph attention network is designed to encode the event relation graph based on hard labels, and update sentence embedding with events and event relations information. The utilization of both soft labels and hard labels of the event relation graph are complementary: soft labels provide nuanced probabilistic information and necessitate the basic language model to recognize that nodes represent events and links represent event relations, while hard labels facilitate updating sentence embeddings with interconnected events and event relation embeddings. Experiments on two benchmark datasets demonstrate the effectiveness of our approach based on the event relation graph, yielding performance gains on both precision and recall. The ablation study confirms the necessity and synergy of leveraging both soft labels and hard labels derived from the event relation graph. Our main contributions are summarized as:

- We firstly observe that interpreting events in association with other events in a document is critical for identifying bias sentences.
- We propose a new framework to incorporate the event relation graph as extra guidance for sentence-level media bias identification.
- We effectively improve the F1 score for bias sentences identification by 5.78% on BASIL dataset and 12.86% on BiasedSents dataset.

2 Related Work

Source-level Media Bias attracted research attention in prior years. The work in this area assumed all the articles and journalists within one media outlet share the same political ideology. Prior work such as Groseclose and Milyo (2005); Gentzkow and Shapiro (2010) measures source-level media bias via citation patterns or consumer political preferences. Budak et al. (2016) implemented a crowdsourcing approach to rate the political slant of media sources. Baly et al. (2018) combined text-based methods with social media behaviors to create a bias classification system.

Article-level Media Bias has been researched for years (Wang, 2017; Lei and Huang, 2023b). Early work used text-based approach (Sapiro-Gheiler, 2019). Chen et al. (2020) developed Gaussian mixture model to incorporate fine-grained bias information. Baly et al. (2020) collected a large-scale dataset and provided manual labels, showing that only three percentage of articles have a different ideology label from their source's. Liu

et al. (2022) pre-trained a political domain language model. Different from previous work on source-level or article-level media bias, we aim to identify media bias at the fine-grained sentence-level, and pinpoint the content that can illuminate and explain the overall article bias.

Sentence-level Media Bias has a relatively short research history (Lei and Huang, 2022, 2023a; Vargas et al., 2023). The initial work in this field is Fan et al. (2019), where they annotate sentence-level media bias within political news document. Another work (Lim et al., 2020) also annotate media bias at sentence level while considering the article context. A discourse structure method was developed by Lei et al. (2022), in which the discourse structures knowledge were distilled into bias sentence identification via a knowledge distillation framework. Different from Lei et al. (2022), our event relation graph interprets news narrative from the perspective of event reporting, and enables constructing event-level discourse relations for every possible pair of sentences, fostering a comprehensive understanding of the broader context.

Event Graph was introduced by Li et al. (2020) where they presented a graph-based structure connecting events and entities nodes through event argument role relations. This event graph was deployed in several down-stream applications, such as sentence fusion (Yuan et al., 2021), story generation (Chen et al., 2021), and misinformation detection (Wu et al., 2022). The event graph constructed in their work comprises entity-entity links and event-entity links via event argument roles, while lacks event-event relations. Differently, our proposed event relation graph focuses on events and establishes event interconnections through four types of event-event relations.

Event and Event Relations were researched for decades. Previous work explored event identification and event extraction (Du and Cardie, 2020; Liu et al., 2020). There are four common relations between events: coreference (Zeng et al., 2020; Barhom et al., 2019), temporal (Ning et al., 2018; Han et al., 2019), causal (Zuo et al., 2021; Gao et al., 2019), and subevent relations (Aldawsari and Finlayson, 2019; Lai et al., 2022). While each type of relation was previously studied in isolation, a recent work (Wang et al., 2022) provided the first large-scale corpus with all four relations annotated. Instead of modeling each event relation separately, we aim to develop a model that unifies all four relations for comprehensive discourse analysis.

3 Event Relation Graph

The event relation graph can incorporate broader contextual information outside the single sentence, and thus help reveal the underlying bias inside the sentence. Hence, we propose to create an event relation graph for each article. In this section, we explain the components, training process, and construction process of the event relation graph.

3.1 Event Relation Graph Components

The event relation graph consists of events as nodes and four types of event relations as links. An event refers to an occurrence or action reported in the article. Coreference relation informs us whether the two events designate the same event. Temporal relation represents the chronological order. Instead of only reflecting the existence of temporal order, we classify temporal relation into three detailed categories for deeper narrative understanding: *before*, *after*, and *overlap*. Causal relation shows the causality or precondition relation between events, and we classify causal relation into two specific types: *causes* or *caused by*. Subevent relation recognizes containment or subordination relation between events, and is categorized into *contains* and *contained by*. Hence, the designed event relation graph not only identifies the presence of each relation, but also offers more detailed and in-depth interconnection information.

3.2 Event Relation Graph Training

The event relation graph is trained on the general-domain MAVEN-ERE dataset (Wang et al., 2022). Following the previous work (Yao et al., 2020), we form the training event pairs in the natural textual order, where the former event in each pair is the precedent event mentioned in the text. In terms of temporal relations, the dataset also annotates the time expressions such as date or time. Considering our event relation graph focuses on events, we solely retain annotations between events. We further process the *before* annotation in this way: keep the *before* label if the annotated event pair aligns with the natural textual order, or reverse the event pair and assign the *after* label if not. The *simultaneous*, *overlap*, *begins-on*, *ends-on*, *contains* annotations are combined into the *overlap* category in our event relation graph. In terms of causal relations, we maintain the *causes* label if the natural textual order is followed, or assign the *caused by* label if not. Similar process for the *contains*

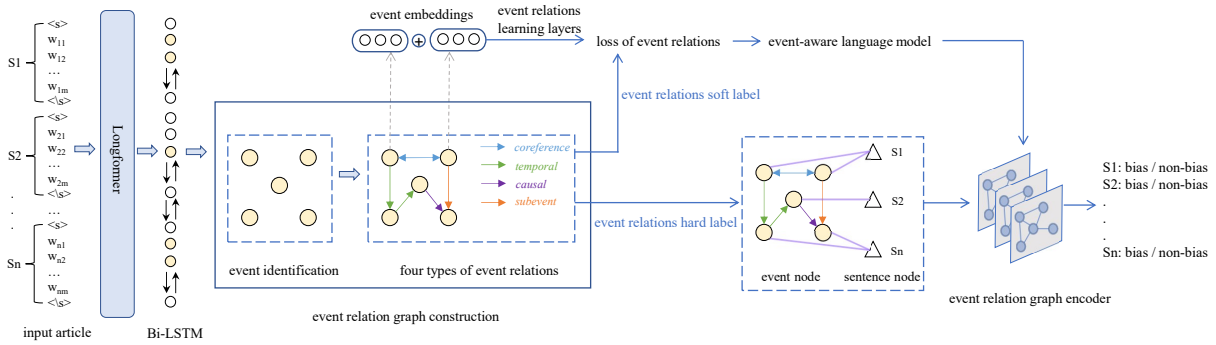


Figure 2: An illustration of sentence-level media bias identification based on event relation graph

and *contained by* labels within subevent relations. Since events and four event relations interact with each other to form a cohesive narrative framework, we adopt joint learning framework (Wang et al., 2022) to train these components collaboratively. The training process results in an event identifier and four event relations extractors.

3.3 Event Relation Graph Construction

Given a candidate news article, an event relation graph is created. During the construction process, we first identify which word triggers an event by using the trained event identifier to generate the predicted probability for each word:

$$P_i^{event} = (p_i^{non-event}, p_i^{event}) \quad (1)$$

where $i = 1, ..., N$, N is the number of words in the article, and p_i^{event} is the probability of i-th word designating an event.

Next, we establish all possible event pairs based on the identified events and extract the four relations between them. To accomplish this, we employ the trained four relations extractors to generate the predicted probabilities for each event pair $(event_i, event_j)$ as follows:

$$P_{i,j}^{corefer} = (p_{i,j}^{non-corefer}, p_{i,j}^{corefer}) \quad (2)$$

$$P_{i,j}^{temp} = (p_{i,j}^{non-temp}, p_{i,j}^{before}, p_{i,j}^{after}, p_{i,j}^{overlap}) \quad (3)$$

$$P_{i,j}^{causal} = (p_{i,j}^{non-causal}, p_{i,j}^{cause}, p_{i,j}^{caused-by}) \quad (4)$$

$$P_{i,j}^{subevent} = (p_{i,j}^{non-subevent}, p_{i,j}^{contain}, p_{i,j}^{contained-by}) \quad (5)$$

where i and j denote the index of two events. $p_{i,j}^{corefer}$ is the probability of two events corefer with each other. $p_{i,j}^{non-temp}$ indicates the probability of no temporal relation existing between them, and $p_{i,j}^{before}, p_{i,j}^{after}, p_{i,j}^{overlap}$ represents the probability of the relations *before*, *after*, *overlap* correspondingly. Similar denotations for the causal and subevent relations.

In the latter bias sentences identification process, the predicted probabilities in equation (1)-(5) are incorporated as the *soft labels* for the constructed event relation graph. The *hard labels* for the events and four relations are derived by applying the $argmax$ function on the soft labels.

4 Bias Sentences Identification

The event relation graph is incorporated into bias sentences identification within two consecutive steps, as illustrated in Figure 2. Firstly, an event-aware language model is developed to inject the events and event relations knowledge into the basic language model using the *soft labels*. Secondly, a relation-aware graph attention network is devised to encode the event relation graph based on *hard labels*, and update sentence embeddings with events and event relations information.

4.1 Event-Aware Language Model

The event-aware language model aims to inject the events and event relations knowledge from the soft labels into the basic language model. Since the news articles are usually long, we utilize the Longformer (Beltagy et al., 2020) as the basic language model to encode the entire article and derive each word embedding. We also add an extra layer of Bi-LSTM (Huang et al., 2015) on top to capture the contextual information.

To enhance the basic language model with the event relation graph knowledge, we construct an event learning layer on top of word embeddings $(w_1, w_2, ..., w_N)$ to learn the events knowledge, and also build four relations learning layers on top of event pair embeddings $(e_i \oplus e_j)$ to learn the four

event relations respectively:

$$Q_i^{event} = (q_i^{non-event}, q_i^{event}) = softmax(W_2(W_1 w_i + b_1) + b_2) \quad (6)$$

$$Q_{i,j}^{corefer} = (q_{i,j}^{non-corefer}, q_{i,j}^{corefer}) = softmax(W_4(W_3(e_i \oplus e_j) + b_3) + b_4) \quad (7)$$

$$Q_{i,j}^{temp} = (q_{i,j}^{non-temp}, q_{i,j}^{before}, q_{i,j}^{after}, q_{i,j}^{overlap}) = softmax(W_6(W_5(e_i \oplus e_j) + b_5) + b_6) \quad (8)$$

$$Q_{i,j}^{causal} = (q_{i,j}^{non-causal}, q_{i,j}^{cause}, q_{i,j}^{caused-by}) = softmax(W_8(W_7(e_i \oplus e_j) + b_7) + b_8) \quad (9)$$

$$Q_{i,j}^{subevent} = (q_{i,j}^{non-subevent}, q_{i,j}^{contain}, q_{i,j}^{contained-by}) = softmax(W_{10}(W_9(e_i \oplus e_j) + b_9) + b_{10}) \quad (10)$$

where $W_1, ..., W_{10}, b_1, ..., b_{10}$ are trainable parameters in the learning layers. $(e_i \oplus e_j)$ is the embedding for the event pair $(event_i, event_j)$ by concatenating the two events embeddings together. Q_i^{event}, and $Q_{i,j}^{corefer}, Q_{i,j}^{temp}, Q_{i,j}^{causal}, Q_{i,j}^{subevent}$ are the learned probabilities for events and four event relations of the basic language model.

The soft labels generated by the event relation graph $P_i^{event}, P_{i,j}^{corefer}, P_{i,j}^{temp}, P_{i,j}^{causal}, P_{i,j}^{subevent}$ contain informative knowledge of events and event relations, thereby are referenced as the target learning materials. By minimizing the cross entropy loss between the learned probability and the target probability, the events and event relations knowledge within the event relation graph can be injected into the basic language model:

$$Loss_{event} = -\sum_{i=1}^{N} P_i^{event} \log(Q_i^{event}) \quad (11)$$

$$Loss_r = -\sum_{i,j} P_{i,j}^r \log(Q_{i,j}^r) \quad (12)$$

where $r \in \{corefer, temp, causal, subevent\}$ represents the four event relations. The overall learning objective for training the event-aware language model based on soft labels is to minimize the cumulative loss for learning each component:

$$Loss_{soft} = Loss_{event} + Loss_{corefer} + Loss_{temp} + Loss_{causal} + Loss_{subevent} \quad (13)$$

4.2 Event Relation Graph Encoder

The event relation graph encoder is designed to encode the event relation graph based on hard labels. In this process, event embeddings are aggregated with neighbor events embeddings through interconnected relations, and sentence embeddings are updated with events and event relations embeddings.

To capture sentence-level context and explicitly demonstrate the connection between each sentence and its reported events, we introduce an extra node to represent each sentence and connect it with the associated events, as depicted in Figure 2.

The resulting event relation graph contains two types of nodes (event nodes and sentence nodes) and nine types of heterogeneous relations (*coreference*, *before*, *after*, *overlap*, *causes*, *caused by*, *contains*, *contained by* relations between events, and event-sentence relation). The event node is initialized as the event word embedding, and the sentence node is initialized as the embedding of the sentence start token <s>. The eight event-event relations are constructed based on hard labels and inherently carry semantic meaning. In order to integrate the semantic meaning of event relations into graph propagation, we introduce the relation-aware graph attention network. In terms of event-sentence relation which is a standard link without semantics, we utilize the standard graph attention network (Veličković et al., 2018).

The relation-aware graph attention network processes event-event relations, and propagates relations semantic meaning into the connected events embeddings. The event-event relation r_{ij} connecting i-th and j-th event nodes is initialized as the embedding of the corresponding relation word r. During the propagation at the l-th layer, the input of i-th event node is the output produced by the previous layer denoted as $e_i^{(l-1)}$, and the relation embedding r_{ij} is updated as:

$$r_{ij} = W^r[e_i^{(l-1)} \oplus r_{ij} \oplus e_j^{(l-1)}] \quad (14)$$

where $\oplus$ represents feature concatenation and W^r is a trainable matrix. Then the attention weights across neighbor event nodes are computed as:

$$\alpha_{ij} = softmax_j\big((W^Q e_i^{(l-1)})(W^K r_{ij})^T\big) \quad (15)$$

where W^Q, W^K are trainable matrices. The output feature for i-th event node regarding the relation type r is formulated as:

$$e_{i,r}^{(l)} = \sum_{j \in \mathcal{N}_{i,r}} \alpha_{ij} W^V r_{ij} \quad (16)$$

where W^V is a parameter, and $\mathcal{N}_{i,r}$ denotes the neighbor event nodes connecting with i-th event via the relation type r. The same procedure to derive the i-th event embedding for all the relation types $r \in R = \{$*coreference*, *before*, *after*, *overlap*, *causes*, *caused*

by, *contains*, *contained by*} . The final output feature for i-th event node at the l-th layer is accumulated as:

$$e_i^{(l)} = \sum_{r \in R} e_{i,r}^{(l)} / |R| \quad (17)$$

The standard graph attention network handles the event-sentence relation, and updates sentence node embeddings with interconnected events and event relations embeddings. During the propagation at l-th layer, the input of k-th sentence node is the output feature produced by the previous layer denoted as $s_k^{(l-1)}$, and the attention weights across the connected events are:

$$h_{kj} = LeakyReLU\left(a^T \left[W s_k^{(l-1)} \oplus W e_j^{(l-1)}\right]\right) \quad (18)$$

$$\alpha_{kj} = softmax_j(h_{kj}) = \frac{\exp(h_{kj})}{\sum_{j \in \mathcal{N}_k} \exp(h_{kj})} \quad (19)$$

where $\mathcal{N}_k$ is the set of event nodes reported in k-th sentence. The final output feature for k-th sentence at the l-th layer is calculated as:

$$s_k^{(l)} = \sum_{j \in \mathcal{N}_k} \alpha_{kj} W e_j^{(l-1)} \quad (20)$$

The sentence node embedding generated at the last layer captures the features of interconnected events and event relations, encompassing both the graph structure and sentence context. We further build a two-layer classification head on top to predict whether the sentence contains bias. The classical cross entropy loss is used for training.

5 Experiments

5.1 Datasets

Among the existing datasets, BASIL (Fan et al., 2019) and BiasedSents (Lim et al., 2020) are the only two datasets that annotate sentence-level media bias while considering the article context. Other studies (Spinde et al., 2021b,a) annotated bias sentences individually without taking into account the broader context. Recognizing that media bias can be very subtle and usually depends on comprehensive contextual information to be discovered (Fan et al., 2019), we utilize BASIL and BiasedSents datasets in the subsequent experiments. Table 1 shows the statistics of the two datasets.

- **BASIL** dataset (Fan et al., 2019) gathered 300 articles from the period of 2010 to 2019. Both lexical bias and informational bias are annotated: lexical bias shifts public opinions through overt ideological language, while informational bias selectively includes or omits information. Because both types of bias can introduce ideological bias to the readers and sway their opinions, we consider them both in our bias sentences identification task. To be specific, we label a sentence as *bias* if it carries either type of bias, or assign the *non-bias* label if neither type of bias exists.

- **BiasedSents** (Lim et al., 2020) collected 46 articles from the year of 2017 to 2018. Each sentence is annotated into four scales: not biased, slightly biased, biased, and very biased. Following the previous work on binary judgments (Fan et al., 2019; Lei et al., 2022), we also process the first two scales as *non-bias* class and the latter two as *bias* class. The dataset releases the annotations from five different annotators, from which we derive the majority voting label as the ground truth.

Dataset	# Article	# Sent	# Bias	% Bias
BASIL	300	7977	1623	20.34
BiasedSents	46	842	290	34.44

Table 1: Number of articles, sentences, bias sentences, and the ratio of bias sentences in the two datasets.

5.2 Evaluation of Event Relation Graph

The event relation graph is trained on the recent MAVEN-ERE dataset (Wang et al., 2022) which annotates all the four event relations within a large scale of general-domain news articles. We employ the current state-of-art model framework (Wang et al., 2022) to train different components collaboratively. Table 4 presents the performance of event identification. Table 5 shows the performance of event coreference resolution. Following the previous work (Cai and Strube, 2010), MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998), $CEAF_e$ (Luo, 2005), and BLANC (RECASENS and HOVY, 2011) are used as evaluation metrics. The performances of other components in the event relation graph, including temporal, causal, and subevent relation classification are summarized in Table 6. The standard macro-average precision, recall, and F1 score are reported.

5.3 Experimental Settings

In the experiments, the model takes the entire news article as input, and outputs the prediction for each

	BASIL			BiasedSents		
	Precision	Recall	F1	Precision	Recall	F1
Baseline Model						
all-bias	20.34	100.00	33.81	34.44	100.00	51.23
gpt-3.5-turbo	32.22	42.39	36.61	42.22	30.25	35.25
gpt-3.5-turbo + 5-shot	33.57	43.07	37.73	42.08	32.16	36.46
gpt-3.5-turbo + full article	37.58	44.45	40.73	40.15	34.39	37.05
gpt-3.5-turbo + event relation graph	32.34	52.86	40.13	40.93	48.06	44.21
(Lei et al., 2022)	49.41	48.80	49.10	42.81	83.44	56.59
longformer	46.81	45.65	46.22	39.78	79.30	52.98
longformer + additional features	47.02	46.33	46.67	40.27	79.65	53.49
Event Relation Graph						
+ event-aware language model (soft label)	50.97	48.61	49.76	46.47	79.62	58.68
+ event relation graph encoder (hard label)	47.28	53.11	50.03	49.72	84.39	62.57
+ both (full model)	**50.06**	**54.10**	**52.00**	**54.10**	**84.08**	**65.84**

Table 2: Performance of sentence-level media bias identification based on event relation graph. Precision, Recall, and F1 score of the *bias* class are shown. The model with the best performance is **bold**.

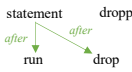
sentence within the article. Follow the previous work (Fan et al., 2019; Lei et al., 2022), we also perform ten-folder cross validation. Instead of splitting the dataset into ten folders based on individual sentences like Fan et al. (2019) did, we follow Lei et al. (2022) to divide the ten folders based on articles. This ensures sentences from the same article do not appear in both the training and testing folds, thus preventing knowledge leakage. In each iteration, one fold is designated as the test set, eight folds are utilized as the training set, and the remaining fold is the validation set to determine the stopping point for training. The maximum training epochs for each iteration is set to 5. We utilize the AdamW optimizer (Loshchilov and Hutter, 2019), with a linear scheduler to adaptively adjust the learning rate. The weight decay is set to 1e-2. Upon collecting the prediction results for the ten testing folds, precision, recall, and F1 score of the *bias* class are calculated for evaluation.

5.4 Baselines

Sentence-level media bias has a relatively short research history, and there are only a few established methods available for comparison. The following systems are implemented as baselines:

- **all-bias**: a naive baseline that predicts all the sentences into the *bias* class
- **gpt-3.5-turbo**: an instruction prompt (Appendix A) is crafted to enable the large language model gpt-3.5-turbo to automatically generate the predicted labels
- **gpt-3.5-turbo + 5-shot**: provide five *bias* class examples and five *non-bias* class examples as demonstrations into the gpt-3.5-turbo prompt
- **gpt-3.5-turbo + full article**: provide the full article context into the gpt-3.5-turbo prompt
- **gpt-3.5-turbo + event relation graph**: guide the large language model gpt-3.5-turbo to reason the event relation graph of the article, and then make the prediction for each sentence through a chain-of-thought process (Wei et al., 2023) (Appendix A)
- Lei et al. (2022): our previous work that incorporates news discourse structures and discourse relations between sentences for bias sentence identification. For fair comparison, we update the language model used in that work into Longformer (Beltagy et al., 2020) and include sentences that contain either lexical bias or informational bias as *bias* sentences in the BASIL dataset. This ensures that both the language model and dataset processing are the same across models.
- **longformer**: the same language model Longformer (Beltagy et al., 2020) is used to encode the article and an extra layer of Bi-LSTM (Huang et al., 2015) is added on top. The hidden state at the sentence start token is extracted as the sentence embedding. This baseline model is equivalent to our developed model without the event relation graph.

successful example

sentences		event relation graph	prediction
S1	Donald J. Trump said in a **statement** on Tuesday that he **dropped** plans to moderate a presidential debate.		
S6	Before this statement, Mr. Trump has **run** for president several times since the 1980s but has always decided not to follow through – he chose to **drop** the debate every time.		P (bias | baseline) = 0.38 P (bias | event relation graph) = 0.74

failing example

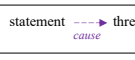
sentences		event relation graph	prediction
S7	Mr. Trump's **statement** seemed certain to open the door to another round of "I just might …" **threats** from the real estate mogul.		P (bias | baseline) = 0.26 P (bias | event relation graph) = 0.48

Figure 3: Example of our method succeed in solving false negative error, and a failing example. Bias sentences are highlighted in red. Events words are shown in **bold** text. The solid arrows in the event relation graphs represent the successfully extracted event relations, and the dashed arrow means the missing event relation.

- **longformer + additional features**: concatenates the probabilistic vector eq (1)-(5) as additional features with the sentence embedding.

5.5 Experimental Results

The experimental results of bias sentences identification based on event relation graph are presented in Table 2. The last row is the performance of our developed model by leveraging both soft labels and hard labels within the event relation graph.

The results demonstrate that incorporating the event relation graph can significantly improve both precision and recall. Compared to the longformer baseline that lacks event relation graph, our proposed method can improve both precision and recall for the BASIL and BiasedSents datasets. This indicates that the event relation graph captures contextual knowledge and facilitates a more comprehensive understanding of the broader discourse.

Moreover, the event relation graph method performs better than the simple feature concatenation. The model that concatenates probabilistic features (longformer + additional features) exhibits only marginal performance improvements over the longformer baseline. This highlights the necessity to develop more sophisticated model to fully harness the information within the event relation graph.

Furthermore, the event relation graph that constructs event-level content structure outperforms the method based on sentence-level relations. Compared to the method incorporating sentence relations (Lei et al., 2022), leveraging the event relation graph leads to superior performance. One possible explanation is that our event relation graph interprets news narratives at a more fine-grained event level, and captures event-level discourse relations across the entire article. This capability enables a more detailed comprehension of each sentence and a deeper understanding of the overall context.

In addition, our method based on event relation graph outperforms the gpt-3.5-turbo baselines. We observe that gpt-3.5-turbo baselines always choose *bias* label when the text is explicitly sentimental. However, the sentence-level ideological bias can be very subtle and usually expressed in a neutral tone (van den Berg and Markert, 2020). This demonstrates that encoding broader context is necessary to discover such implicit ideological bias. Besides, we observe that guiding the large language model to reason the event relation graph (gpt-3.5-turbo + event relation graph) significantly increases the Recall of bias sentences. By leveraging the knowledge from the event relation graph, the large language model gains better understanding of the article context, thereby uncovering more implicit bias cases.

	Precision	Recall	F1
baseline	46.81	45.65	46.22
full model	**50.06**	**54.10**	**52.00**
- coreference	48.36	53.48	50.79
- temporal	48.56	52.98	50.68
- causal	49.91	50.71	50.30
- subevent	48.65	52.37	50.45

Table 3: Effect of removing each of the four event relations. Take BASIL dataset as an example.

5.6 Ablation Study

The ablation study of leveraging either soft labels or hard labels are reported in Table 2. The utilization of soft labels to train an event-aware language model enhances performances on both datasets, especially precision. This indicates that soft labels guide the model to learn nuanced probabilistic information from the event relation graph, enabling it to make more precise predictions. The employment of hard labels to build a heterogeneous graph attention network also improves metrics on both datasets, especially recall. This implies that hard labels facilitate propagating sentence embeddings with interconnected events embeddings,

thereby encourage the encoding of a broader contextual knowledge and enable recalling more implicit cases. Incorporating them both exhibit the best performance, demonstrating that soft and hard labels are complementary with each other.

5.7 Effect of Four Event Relations

We further study the effect of four event relations. The experimental results of removing each of the four event relations from the full model are shown in Table 3. The results indicate that removing any one of the four relations leads to a performance decrease in both precision and recall. The four commonly-seen event relations are essential in constructing a unified and cohesive content structure. Incorporating all the four relations in the event relation graph yields the best performance.

5.8 Analysis and Discussion

Figure 3 shows a successful example in which the event relation graph effectively recognizes the temporal relation. As a result, the model is aware that the sentence (S6) intends to disclose the past habitual behaviors of Trump, and successfully recalls it as *bias* class. The current trained event relation graph has the ability to extract most easy cases through explicit discourse connectives such as *before* or *since* in this example.

Figure 3 also shows a failing example, where the event relation graph fails to recognize the causal relation and leads to a false negative error. It is because the text expresses the causal relation in an implicit way by using the phrase *open the door to*. Such implicit cases that lack explicit cues pose challenges for the current event relation graph. To further improve the performance of bias sentences identification, it is necessary to enhance the extraction of implicit event relations.

6 Conclusion

This paper identifies media bias at the sentence level, and demonstrates the pivotal role of events and event-event relations in identifying bias sentences. We observe that interpreting events in association with other events in a document is critical in identifying bias sentences. Inspired by this observation, this paper proposes to construct an event relation graph as the extra guidance for bias sentence detection. To integrate the event relation graph, we design a novel framework that consists of two consecutive steps to first update the language model and then further update sentence embeddings, by leveraging the soft and hard event relation labels respectively from the event relation graph. Experimental results demonstrate the effectiveness of our approach. For future work, we will further improve the extraction of implicit event relations as well as develop new interpretable methods that leverage event relation graph for media bias analysis.

Limitations

Our paper propose to construct an event relation graph for each article to facilitate ideological bias identification. The current state-of-art model framework is employed for training the event relation graph. While the current event relation graph successfully extracts event relations for most easy cases with explicit discourse connectives or language cues, it may encounter challenges in recognizing implicitly stated event relations. Thus, improving event relation graph construction becomes necessary and serves as the future work.

Ethical Considerations

This paper is a research paper on identifying media bias. The goal of this research is to understand, detect, and mitigate the political bias. The examples in this paper are only used for research purpose, and do not represent any political leaning of the authors. The release of the datasets and code should be used for mitigating the political bias, instead of expanding or disseminating the political bias.

Acknowledgements

We would like to thank the anonymous reviewers for their valuable feedback and input. We gratefully acknowledge support from National Science Foundation via the awards IIS-1942918 and IIS-2127746. Portions of this research were conducted with the advanced computing resources provided by Texas A&M High-Performance Research Computing.

References

Mohammed Aldawsari and Mark Finlayson. 2019. Detecting subevents using discourse and narrative features. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4780–4790, Florence, Italy. Association for Computational Linguistics.

Amit Bagga and Breck Baldwin. 1998. Entity-based cross-document coreferencing using the vector space

model. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL '98/COLING '98, page 79–85, USA. Association for Computational Linguistics.

Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. 2020. 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.

Shany Barhom, Vered Shwartz, Alon Eirew, Michael Bugert, Nils Reimers, and Ido Dagan. 2019. Revisiting joint modeling of cross-document entity and event coreference resolution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4179–4189, Florence, Italy. Association for Computational Linguistics.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer.

Ceren Budak et al. 2016. Fair and balanced? quantifying media bias through crowdsourced content analysis. *Public Opinion Quarterly*, 80(S1):250–271.

Jie Cai and Michael Strube. 2010. Evaluation metrics for end-to-end coreference resolution systems. In *Proceedings of the SIGDIAL 2010 Conference*, pages 28–36, Tokyo, Japan. Association for Computational Linguistics.

Hong Chen, Raphael Shu, Hiroya Takamura, and Hideki Nakayama. 2021. GraphPlan: Story generation by planning with event graph. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 377–386, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Wei-Fan Chen, Khalid Al Khatib, Benno Stein, and Henning Wachsmuth. 2020. Detecting media bias in news articles using Gaussian bias distributions. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4290–4300, Online. Association for Computational Linguistics.

Xinya Du and Claire Cardie. 2020. Event extraction by answering (almost) natural questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 671–683, Online. Association for Computational Linguistics.

Robert M. Entman. 2007. 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.

Lei Gao, Prafulla Kumar Choubey, and Ruihong Huang. 2019. Modeling document-level causal structures for event causal relation identification. 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 1808–1817, Minneapolis, Minnesota. Association for Computational Linguistics.

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

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

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

Tim Groseclose and Jeffrey Milyo. 2005. A measure of media bias. *The Quarterly Journal of Economics*, 120(4):1191–1237.

Rujun Han, Qiang Ning, and Nanyun Peng. 2019. Joint event and temporal relation extraction with shared representations and structured prediction. 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 434–444, Hong Kong, China. Association for Computational Linguistics.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging.

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.

Viet Lai, Hieu Man, Linh Ngo, Franck Dernoncourt, and Thien Nguyen. 2022. Multilingual SubEvent relation extraction: A novel dataset and structure induction method. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5559–5570, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yuanyuan Lei and Houwei Cao. 2023. Audio-visual emotion recognition with preference learning based on intended and multi-modal perceived labels. *IEEE*

Transactions on Affective Computing, 14(4):2954–2969.

Yuanyuan Lei and Ruihong Huang. 2022. Few-shot (dis)agreement identification in online discussions with regularized and augmented meta-learning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5581–5593, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yuanyuan Lei and Ruihong Huang. 2023a. Discourse structures guided fine-grained propaganda identification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 331–342, Singapore. Association for Computational Linguistics.

Yuanyuan Lei and Ruihong Huang. 2023b. Identifying conspiracy theories news based on event relation graph. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9811–9822, Singapore. Association for Computational Linguistics.

Yuanyuan Lei, Ruihong Huang, Lu Wang, and Nick Beauchamp. 2022. Sentence-level media bias analysis informed by discourse structures. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10040–10050, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Manling Li, Qi Zeng, Ying Lin, Kyunghyun Cho, Heng Ji, Jonathan May, Nathanael Chambers, and Clare Voss. 2020. Connecting the dots: Event graph schema induction with path language modeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 684–695, Online. Association for Computational Linguistics.

S Robert Lichter. 2017. Theories of media bias. In *The Oxford handbook of political communication*, page 403. Oxford University Press New York.

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

Jian Liu, Yubo Chen, Kang Liu, Wei Bi, and Xiaojiang Liu. 2020. Event extraction as machine reading comprehension. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1641–1651, Online. Association for Computational Linguistics.

Yujian Liu, Xinliang Frederick Zhang, David Wegsman, Nicholas Beauchamp, and Lu Wang. 2022. POLITICS: Pretraining with same-story article comparison for ideology prediction and stance detection. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1354–1374, Seattle, United States. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 25–32, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

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

Qiang Ning, Hao Wu, and Dan Roth. 2018. A multi-axis annotation scheme for event temporal relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1318–1328, Melbourne, Australia. Association for Computational Linguistics.

Markus Prior. 2013. Media and political polarization. *Annual Review of Political Science*, 16:101–127.

M. RECASENS and E. HOVY. 2011. Blanc: Implementing the rand index for coreference evaluation. *Natural Language Engineering*, 17(4):485–510.

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 Press.

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. *CoRR*, abs/2105.11910.

Jesper Strömbäck. 2012. The media and their use of opinion polls: Reflecting and shaping public opinion. *Opinion polls and the media: Reflecting and shaping public opinion*, pages 1–22.

Esther van den Berg and Katja Markert. 2020. Context in informational bias detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6315–6326, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Teun A Van Dijk et al. 1995. Power and the news media. *Political communication and action*, 6(1):9–36.

James K Van Leuven and Michael D Slater. 1991. How publics, public relations, and the media shape the public opinion process. *Journal of Public Relations Research*, 3(1-4):165–178.

Francielle Vargas, Kokil Jaidka, Thiago Pardo, and Fabrício Benevenuto. 2023. Predicting sentence-level factuality of news and bias of media outlets. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 1197–1206, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

Xiaozhi Wang, Yulin Chen, Ning Ding, Hao Peng, Zimu Wang, Yankai Lin, Xu Han, Lei Hou, Juanzi Li, Zhiyuan Liu, Peng Li, and Jie Zhou. 2022. MAVEN-ERE: A unified large-scale dataset for event coreference, temporal, causal, and subevent relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 926–941, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Anne E Wilson, Victoria A Parker, and Matthew Feinberg. 2020. Polarization in the contemporary political and media landscape. *Current Opinion in Behavioral Sciences*, 34:223–228.

Xueqing Wu, Kung-Hsiang Huang, Yi Fung, and Heng Ji. 2022. Cross-document misinformation detection based on event graph reasoning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 543–558, Seattle, United States. Association for Computational Linguistics.

Wenlin Yao, Zeyu Dai, Maitreyi Ramaswamy, Bonan Min, and Ruihong Huang. 2020. Weakly Supervised Subevent Knowledge Acquisition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5345–5356, Online. Association for Computational Linguistics.

Ruifeng Yuan, Zili Wang, and Wenjie Li. 2021. Event graph based sentence fusion. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4075–4084, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yutao Zeng, Xiaolong Jin, Saiping Guan, Jiafeng Guo, and Xueqi Cheng. 2020. Event coreference resolution with their paraphrases and argument-aware embeddings. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3084–3094, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Xinyu Zuo, Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Weihua Peng, and Yuguang Chen. 2021. Improving event causality identification via self-supervised representation learning on external causal statement. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2162–2172, Online. Association for Computational Linguistics.

A Prompt for gpt-3.5-turbo baselines

The instruction prompt for gpt-3.5-turbo baseline that takes individual sentences as input is: "Biased sentence with political bias refers to the supportive or background content to sway readers opinions in an ideological direction, either through information selective inclusion or overt ideological language. Please reply "Yes" if the following sentence is a biased sentence with political bias, else reply "No". Sentence: "<sentence>" Answer:"

The instruction prompt for gpt-3.5-turbo + full article baseline that takes the full article context as input is: "Biased sentence with political bias refers to the supportive or background content to sway readers opinions in an ideological direction, either through information selective inclusion or overt ideological language. Given the article: "<article>". Please reply "Yes" if the following sentence is a biased sentence with political bias, else reply "No". Sentence: "<sentence>" Answer:"

The instruction prompt for gpt-3.5-turbo + event relation graph baseline that guides the model to reason the event relation graph of the article and then make the prediction for each sentence through a chain-of-through process is: "Biased sentence with political bias refers to the supportive or background content to sway readers opinions in an ideological direction, either through information selective inclusion or overt ideological language. The task is identifying whether a sentence is biased or not. Let's think step by step. Firstly, explain the events reported in the sentence and the event relations

in the article context. Events refer to an occurrence or action reported in the sentence. There are four types of relations between events: coreference, temporal, causal, and subevent relations. Coreference relation represents whether two event mentions designate the same occurrence. Temporal relation represents the chronological order between events, such as before, after, and overlap. Causal relation represents the causality between events. Subevent relation represents the containment relation from a parent event to a child event. Secondly, answer "Yes" if the following sentence is a biased sentence with political bias, else reply "No". Please mimic the output style in the following example.
Article: "Donald J. Trump said in a statement on Tuesday that he dropped plans to moderate a presidential debate. Trump said that he dropped because the Republican Party asked him to give up the right to run as an independent candidate if he moderates the debate. "It is very important to me that the right Republican candidate be chosen to defeat the failed Obama administration," Mr. Trump said in his statement. "But if that Republican, in my opinion, is not the right candidate, I am not willing to give up my right to run as an independent candidate." Mr. Trump drew considerable criticism from the Republican establishment and attracted the ridicule from the public for his decision to drop and his willingness to still run if he was displeased with the candidates. Before this statement, Mr. Trump has run for president several times since the 1980s but has always decided not to follow through – he chose to drop the debate every time."
Sentence: "Before this statement, Mr. Trump has run for president several times since the 1980s but has always decided not to follow through – he chose to drop the debate every time."
Output: Firstly, explain the events reported in the sentence and the event relations in the article context. This sentence reports the events: run for president, drop the debate. The event relations in the article context: the event "run for president" is a historical event that happened before the event "statement". Trump claimed a causal relation between the event "drop the debate" and the event "give up right as an independent candidate". This sentence reports a historical event to reveal Trump's consistent past behavior of dropping presidential election several times since 1980s, and hints that the claimed cause by Trump is not his real reason to drop the debate. Therefore, this sentence contains political bias to sway readers' opinion towards Trump to a negative direction. Secondly, answer "Yes" if the following sentence is a biased sentence with political bias, else reply "No". Answer: Yes.
Article: "<article>"
Sentence: "<sentence>"
Output:"

B Evaluation of Event Relation Graph

This section reports the performance of the event relation graph. The event relation graph is trained on the recent MAVEN-ERE dataset (Wang et al., 2022) which annotates all the four event relations within a large scale of general-domain news articles. Since events and four event relations interact with each other to form a cohesive narrative framework, we adopt joint learning framework (Wang et al., 2022) to train different components collaboratively. The AdamW (Loshchilov and Hutter, 2019) is used as the optimizer. The training epochs is 10. The learning rate is initialized as 1e-5 and adjusted by a linear scheduler. The weight decay is set to 1e-2.

Table 4 presents the performance of event identification. Table 5 shows the performance of event coreference resolution. Following the previous work (Cai and Strube, 2010), MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998), $CEAF_e$ (Luo, 2005), and BLANC (RECASENS and HOVY, 2011) are used as evaluation metrics. The performances of other components in the event relation graph, including temporal, causal, and subevent relation classification are summarized in Table 6. The standard macro-average precision, recall, and F1 score are reported.

	Precision	Recall	F1
Event Identifier	87.31	91.81	89.40

Table 4: Performance of event identification. Macro precision, recall, and F1 are reported.

MUC			B^3			$CEAF_e$			BLANC		
Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
76.34	83.10	79.57	97.07	98.32	97.69	97.79	97.00	97.39	83.69	92.43	87.54

Table 5: Performance of event coreference resolution in the event relation graph

	Precision	Recall	F1
Temporal	48.45	46.43	47.04
Causal	58.48	54.02	56.01
Subevent	53.37	42.90	46.21

Table 6: Performance of temporal, causal, and subevent relation tasks in the event relation graph. Macro precision, recall, and F1 are reported.