All Things Considered: Detecting Partisan Events from News Media with Cross-Article Comparison

Anonymous ACL submission

Abstract

Public opinion is shaped by the information news media provide, and that information in turn may be shaped by the ideological preferences of media outlets. But while much attention has been devoted to media bias via overt ideological language or topic selection, a more profound way in which the media shape opinion is via the strategic inclusion or omission of *partisan events* that may support one side or the other. We develop a latent variable-based framework to predict the ideology of news articles by comparing multiple articles on the same story and identifying partisan events whose inclusion or omission reveals ideology. Our experiments first validate the existence of partisan event selection, and then show that article alignment and cross-document comparison detect partisan events and article ideology better than existing methods. Our results reveal the high-level form of media bias, which is present even among mainstream media with strong norms of objectivity and nonpartisanship.

1 Introduction

News media play a critical role in society not merely by supplying information, but also by selecting and shaping the content they report (de Vreese, 2004; DellaVigna and Kaplan, 2007; DellaVigna and Gentzkow, 2009; Perse and Lambe, 2016). To understand how media bias affects media consumers (Gentzkow and Shapiro, 2006; Gentzkow et al., 2015), we must understand not just how media ideology affects the presentation of news stories on a surface level, such as the usage of partisan phrases or opinions, but also the less obvious process of content selection (Fan et al., 2019; Enke, 2020). Content selection, such as what events that are related to the main story and should be included in the report, has recently become a focus of study in political science. Numerous studies point out that media selectively report information that is flattering to a particular political party or ideology, which may consequently shift audience beliefs and attitudes (Broockman and Kalla, 2022; Baum and Groeling, 2008; Grossman et al., 2022; D'Alessio and Allen, 2006). However, most existing work either requires manual inspection of reported content (Broockman and Kalla, 2022), or relies on simple tools for coarse analyses, such as overall slant and topic emphasis (Baum and Groeling, 2008; Grossman et al., 2022). As a result, these studies are either limited to a short time period, or are unable to provide a detailed understanding of content selection bias. Thus there remains a strong need for automatic tools that can analyze and detect how more complex content is selectively reported.

News Story: *Biden pushes for gun legislation after visiting Uvalde.*

The Washington Post (left):
E1: [Jaydien]$_{\text{ARG0}}$, . . ., [said]$_{\text{pred}}$ [he asked the president: "Could you please make our schools safer and send more police, please?"]$_{\text{ARG1}}$
E2: [Biden]$_{\text{ARG0}}$ . . . [noting]$_{\text{pred}}$: "[You couldn't buy a cannon when the Second Amendment was passed]$_{\text{ARG1}}$."

New York Post (right):
E1: [You]$_{\text{ARG0}}$ couldn't [buy]$_{\text{pred}}$ [a cannon]$_{\text{ARG1}}$ when the Second Amendment was passed.
E2: Biden has made that claim before, . . ., and they have been repeatedly [declared]$_{\text{pred}}$ [false]$_{\text{ARG1}}$ [by fact-checkers]$_{\text{ARG0}}$.

Figure 1: Article snippets by different media on the same story. Events are represented by triplets of ⟨ARG0, predicate, ARG1⟩. Events favoring left and right sides are highlighted in blue and red. Events in black are reported by both media and not considered as partisan.

Rather than focusing on more superficial biases such as word, topic, or entity selection, we investigate here how media ideology affects their selection of which **events** to include for news reporting. Events are the fundamental high-level components of the storytelling process (Prince, 2012), and their inclusion or omission shapes how a news story is perceived. In line with previous analysis of partisan selection bias in the literature (Broockman and Kalla, 2022), we define **partisan events** as *selectively reported events that are flattering to a*

media organization's copartisans or unfavorable to opponents. When there are many potentially relevant events, which subset are included in an article fundamentally affects how readers interpret the story, and can reveal a media outlet's stance on that topic and their ideology (Mullainathan and Shleifer, 2005; McCombs and Reynolds, 2008; Entman, 2007). One example of event-selection bias is shown in Fig. 1, where a Washington Post article includes a survivor's request to impose gun control (*pro*-gun control), whereas a New York Post article claims Biden's statement as false (*pro*-gun rights).

This paper has two major goals: (1) examining the *relation between event selection and media ideology*, and (2) formulating a task for *partisan event detection in news articles* and developing computational methods to automate the process. For the first goal, we verify the existence of partisan event selection by measuring how event selection affects the performance of media ideology prediction. Specifically, we represent articles using triplets of ⟨ARG0, predicate, ARG1⟩, denoting the set of events they report with participating entities (e.g., in Fig. 1). This representation is shown to be effective in narrative understanding (Chambers and Jurafsky, 2008; Mostafazadeh et al., 2016). We conduct *two studies*. First, we compare article-level ideology prediction performance by using events within a single article vs. contextualizing them with events in other news articles on the same story but as reported by media with different ideologies, inspired by the observation that biased content should be evaluated against other media (Larcinese et al., 2011). We show that the latter setup yields significantly higher F1 scores, suggesting that *cross-article comparison* can identify partisan events and thereby produce more accurate ideology prediction. Second, we *annotate an evaluation* dataset of 50 articles that focus on two recent political issues, where in total we manually label 828 *partisan events* on 1867 sentences. Using this dataset, we show that removing partisan events from the articles hurts ideology prediction performance significantly more than removing similar amounts of randomly selected events.

For the second goal, the most critical challenge in developing computational tools to identify partisan events is the lack of annotation, where manually labeling a large-scale dataset requires domain expertise and is highly time-consuming. For that reason, we use latent variables to represent whether an event is partisan or not, and propose to jointly infer partisan events and predict an article's ideology. Models are trained using article-level ideology labels only, which are easier to obtain. We compare two approaches (Chen et al., 2018; Yu et al., 2019) to train latent variable models and explore two methods for further improvement: (1) steering the model toward events that are selected only by one side, which are more likely to be partisan, and (2) providing prior ideology knowledge with pretrained event representations.

We conduct experiments on two existing news article datasets (Liu et al., 2022; Fan et al., 2019) and our *evaluation only* data with newly annotated partisan events.[1] Results indicate that latent variable models outperform all competitive baselines on both partisan event detection and ideology prediction, where cross-article event comparison is shown to be critical for both tasks. Analysis of the extracted partisan events reveals key challenges in detecting implicit nuanced sentiments and discerning event relations (e.g., main vs. background events), suggesting future research directions.

To the best of our knowledge, this is the first time that computational methods are developed for studying media bias at the event selection level. It is also the first time that automatic models are investigated to detect partisan events. Our results provide new insights into a high-level form of media bias that may be present even in apparently nonpartisan news, enabling a new understanding of how news media content is produced and shaped.

2 Related Work

Media Bias Understanding and Detection. Research in political science, economics, and communication has extensively demonstrated the relationship between news media and ideological bias (Mullainathan and Shleifer, 2005; Gentzkow et al., 2014). According to Broockman and Kalla (2022), there are three common strategies news media use to affect readers: *Agenda setting* (McCombs and Shaw, 1972) refers to when the public's perception of a topic's overall significance is shaped by the amount of news coverage spent on that topic (Field et al., 2018; Grimmer, 2010; Quinn et al., 2010; Kim et al., 2014). *Framing* concerns how media highlight some aspects of the same reality to make them more salient to the public (Entman, 1993; Tsur et al., 2015; Baumer et al., 2015; Card et al., 2015; Liu et al., 2019a). Finally,

[1] Our data and code will be made publicly available.

partisan coverage filtering is used by media to selectively report content that is flattering to their co-partisans or unflattering to opponents. While there is a certain amount of conceptual overlap among these three categories, this work focuses primarily on the third: the selection of which events relevant to the main stories to report, and how that reveals a media outlet's ideology and stance. Compared to previous work in agenda setting, which mainly focuses on the topics of news articles (Field et al., 2018; Kim et al., 2014), our partisan event study focuses on a more thoughtful process for information filtering. Event selection is also subtler than framing, since framing examines how a perspective is evoked through particular phrases (Card et al., 2015), whereas partisan event detection requires both event extraction and cross-article comparison.

While partisan coverage filtering has been studied in political science, detecting it requires human efforts to review all news content (Broockman and Kalla, 2022; Baum and Groeling, 2008), making these methods unscalable and only applicable to short time periods. Grossman et al. (2022) automate the process, but use predefined lists of phrases and simple topic models to determine the overall slant and topic of a news report, which cannot capture more tactful content selection like events. By contrast, we automatically detect partisan events in news articles with methods that are scalable to large quantities of news and operate with *more nuanced factual details* than phrases and topics.

Another line of work that is similar to ours is the detection of *informational bias* (Fan et al., 2019; van den Berg and Markert, 2020), defined as "tangential, speculative, or background information that sways readers' opinions" (Fan et al., 2019). Our work differs in two important aspects: First, their "informational bias" can occur in any text span, and detecting speculative information often requires complex inference and also depends on specific wording. By contrast, by focusing on the presence or absence of events, we target concrete units of potentially partisan information, which can be more easily validated and understood by readers. Second, they train supervised models on annotated biased content, while our latent variable models do not need any labels on partisan events.

Ideology Prediction with Text. Many computational models have been developed to predict ideology using textual data (Gentzkow and Shapiro, 2010; Gerrish and Blei, 2011; Ahmed and Xing, 2010; Nguyen et al., 2013). Recent work, for instance, leverages neural networks to incorporate phrase-level ideology (Iyyer et al., 2014), external knowledge from social media (Kulkarni et al., 2018; Li and Goldwasser, 2019), and large-scale language model pretraining (Liu et al., 2022; Baly et al., 2020). However, most of this computational work focuses directly on ideology prediction, with little attention to the higher-level processes underlying media bias. In particular, ideology prediction may fail for many mainstream media outlets who eschew overtly ideological language, and instead may bias readers only via a more sophisticated information selection procedure at the event level. We demonstrate that incorporating story-level context enables global content comparison over political spectrum, and benefits both partisan event detection and ideology prediction.

3 Event Selection Effect Study

In this section, we verify the *existence of partisan event selection* by examining partisan events' influence on ideology prediction. Using events extracted from articles (§3.1), we design a model that predicts ideology with single- or multi-article context (§3.2), based on the assumption that comparing events included by different media may reveal their ideological leanings. We then manually annotate a dataset with partisan events in news stories (§3.3). Using this dataset and two existing corpora, we show that cross-article content comparison can reveal potential partisan events and removing partisan events hurts ideology prediction (§3.4).

3.1 Event Extraction

We follow the scheme in TimeML which defines events as "situations that happen or occur" (Pustejovsky et al., 2003). We train an event extraction model on the MATRES data (Ning et al., 2018), as its event annotation is not limited to predefined event types, and thus is applicable to the open domain scenario. We use RoBERTa-large (Liu et al., 2019b) that predicts a binary label for each word, deciding whether the word is an event predicate or not. On the same train and test split, our model achieves an F1 score of 89.53, which is on par with the state-of-the-art performance of 90.5 F1 score (Zhang et al., 2021). See Appendix B.1 for details. As verbs and nouns account for 96.8% of event predicates in MATRES dataset, we extract arguments 0 and 1 for verb and noun predicates us-

ing semantic role labeling tools (Shi and Lin, 2019; Gardner et al., 2018),[2] and we only keep predicates that match our event extraction results.

Multiple events can exist in one sentence with overlapping predicates and arguments. We hence remove the shorter event if there is an overlap, as we find that shorter events tend to be less informative. For example, it is easier to determine the partisanship of the event "the leak of a draft opinion would mark a stunning betrayal of the Court's process" than a shorter one on "the leak of a draft opinion." Therefore, we remove an event if its predicate is covered by another event's arguments.

3.2 Ideology Prediction with Events

We build on the narrative embedding model in Wilner et al. (2021) and extend it to include story level context by adding article segment, event frequency, and event position embeddings. This allows us to gauge the effect of partisan events' presence or absence on ideology prediction.

Given N input articles $a_1, \ldots, a_N$ that report on the same news story, §3.1 extracts events in article a_i as $x_1^{(i)}, \ldots, x_{L_i}^{(i)}$, where L_i is the number of events in article a_i. We first use a DistilRoBERTa model to get the embedding $\mathbf{e}$ for an event (Sanh et al., 2019): We input the sentence that contains the event to DistilRoBERTa and get the embeddings $\mathbf{e}_{pred}, \mathbf{e}_{arg0}, \mathbf{e}_{arg1}$ for predicate, ARG0, and ARG1 by taking the average of last-layer token embeddings. If a sentence has multiple events, we mask out other events' tokens when encoding one event, so that the information in one event does not leak to others. We then get $\mathbf{e} = \mathbf{W}[\mathbf{e}_{pred}; \mathbf{e}_{arg0}; \mathbf{e}_{arg1}]$, where ; means concatenation and $\mathbf{W}$ is learnable.[3] We then input all events in one article or all articles on the same story to another transformer encoder (Vaswani et al., 2017) to get contextualized $\mathbf{c}$ for each event:

$$[\mathbf{c}_1^{(1)}, \ldots, \mathbf{c}_{L_N}^{(N)}] = \texttt{Transformer}\left([\mathbf{e}_1^{(1)}, \ldots, \mathbf{e}_{L_N}^{(N)}] + E\right) \tag{1}$$

where `Transformer` is a standard transformer encoder trained from scratch (details in Appendix B.2) and E contains three types of embeddings: **Article embeddings** distinguish the source by associating the index of the article with its events, with a maximum of three articles per story. **Frequency embeddings** highlight the prevalence of events by signaling if an event appears in only one article, more than one but not all articles, or all articles that report the same story. We train one embedding for each category and use lexical matching to determine common events. Finally, **position embeddings** represent the relative position of an event in the article, e.g., partisan events may appear later in the reports. All embeddings are learnable (details in Appendix B.2). Note that Eq. 1 describes the model with story level context as it includes all events in all articles. We also experiment with models that only use events in one article. Finally, the model predicts article's ideology using average representation of all events in the article.

3.3 Partisan Event Dataset Annotation

Since there is no dataset with partisan event annotations for news articles, we manually label a **Partisan Event** (**PEvent**) dataset with 50 articles (1867 sentences) covering two recent controversial issues in the U.S. in 2022: a mass shooting in Texas, and the overturn of *Roe v. Wade*. Since labeling partisan events is costly, which requires both domain knowledge and news annotation experience, we only focus on two broad high-profile topics where the partisanship of all constituent events is already known to coders experienced with US politics. Note that the event dataset contains articles from a separate and **later** time than the training data, and is used for **evaluation purposes** only.

We collect articles from AllSides Headline Roundups section,[4] where groups of three articles that report the same news story are carefully selected by editors to demonstrate "how opposite sides of the media are discussing or framing a subject". For each story, we discard the center ideology article to focus on partisan media coverage. The remaining two articles, together with extracted events, are provided to two college students who have prior news article annotation experience and have gone through careful training of the annotation tasks. They are instructed to first label article ideology,[5] and then partisan events. During annotation, we only annotate left partisan events for left articles and vice versa. Finally, a third annotator compares the annotations and resolves conflicts. Appendix C contains the full annotation guideline.

In total, **828 partisan events** are annotated out of 3035 events detected by our tool. Inter-annotator

[2] `github.com/CogComp/SRL-English` for nouns.

[3] We use a zero vector if ARG0 or ARG1 does not exist.

[4] `https://www.allsides.com/blog/how-does-allsides-create-balanced-news`.

[5] We intentionally annotate articles' ideology rather than using media-level ideology to ensure accurate ideology labels.

	AllSides	Basil	PEvent (ours)
# stories	2,221	67	25
# articles	5,361	134	50
# events detected per article	66.82	48.71	60.70

Table 1: Statistics for AllSides training set, Basil, and PartisanEvent. AllSides **test** set contains 1,416 articles.

agreement calculated using Cohen's κ (Cohen, 1960) is 0.83 for article-level ideology. For partisan event labeling, two annotators achieve $\kappa = 0.43$. On average, 16.56 (27.28%) events are annotated as partisan events per article. Among all partisan events reported by left-leaning media, 98.41% are chosen only by the left side, and 95.09% for the right media. We further check where partisan events are included in the articles, and find that they occur more frequently in the later parts of articles written by right-leaning media (displayed in Fig. 4 in the Appendix). These findings validate our design in §3.2.

3.4 Results for Ideology Prediction

We first compare ideology prediction performance using different model variants in §3.2 and then pick two to study the effect of removing partisan events.

Effects of Cross-Article Event Comparison. We train models on AllSides dataset collected in Liu et al. (2022), where media outlets' ideology is used as articles' ideology. We use articles before 2020 (inclusive) as training and dev data and articles after 2020 (exclusive) as test data. We also evaluate models on Basil (Fan et al., 2019), where ideology is manually annotated similar to §3.3. Likewise, we remove articles of the center ideology. Table 1 presents the statistics for datasets used in this study.

We experiment with multi- and single-article variants of the model, depending on whether the transformer in Eq. 1 has access to events in all or one article. As shown in Table 2, multi-article models that allow content comparison across articles written by different media significantly outperform single-article models, demonstrating the benefits of adding story-level context to reveal partisan events that improve ideology prediction. Among multi-article models, article embeddings lead to the largest gain since it supports cross-article comparison. For experiments in the rest of this paper, we add position embedding for single-article models and all three embeddings for multi-article models.

Effects of Removing Partisan Events. Next, we investigate how would removing partisan events affect models' prediction on ideology. Intuitively,

	AllSides	Basil	PEvent
Single-article	64.10 ± 3.51	55.08 ± 6.01	44.37 ± 2.60
$+pos.$	64.37 ± 0.75	54.78 ± 2.38	45.77 ± 3.46
Multi-article	79.52 ± 1.52	64.91 ± 1.78	76.64 ± 3.16
$+art.$	$\underline{88.61} \pm 0.84$	67.30 ± 2.45	$\mathbf{85.19} \pm 2.28$
$+art. + fre.$	$\mathbf{88.64} \pm 0.56$	$\underline{68.05} \pm 1.33$	$\underline{83.60} \pm 1.67$
$+art. + fre. + pos.$	88.49 ± 0.74	$\mathbf{68.50} \pm 2.07$	83.59 ± 1.67

Table 2: Macro F1 scores for article ideology prediction (average of 5 runs). **Best** results are in bold and second best are underlined. $art.$, $fre.$, and $pos.$ refer to article, frequency, and position embeddings in §3.2.

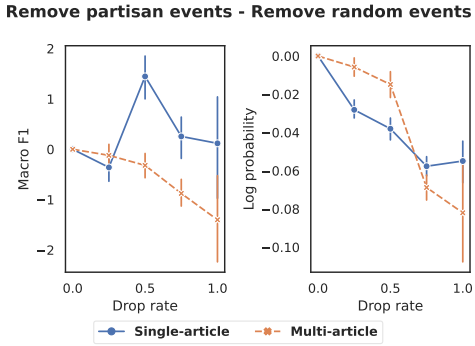

Figure 2: Performance difference (average of 10 runs) between removal of partisan events and of random events. A negative value indicates more severe performance regression when dropping partisan events, compared with dropping the same amount of random events.

when having access to fewer partisan events, models will be less confident in identifying correct ideologies. Concretely, we run experiments on PEvent, by dropping $m\%$ of **partisan events**, where $m = 25, 50, 75, 100$. We also run the same models and remove the same number of events randomly (**random events**). We then measure the performance difference between the two models with macro F1 and log probability of true classes.

As shown in Fig. 2, for multi-article model, removing partisan events hurts the performance more compared to removing random events, indicated by the negative values on y axes. Moreover, the more partisan events are removed, the larger the difference is, which confirms that models exploit the presence of partisan events to discern ideology. However, we observe a different trend of F1 by using the single-article model, indicating possible overfitting of the training set. This again demonstrates the importance of cross-article event comparison.

4 Latent Variable Models for Partisan Event Detection

The general idea of our latent variable models for partisan event detection is that the detected partisan events should be indicative of article's ideology, the

removal of which would lower models' prediction confidence, according to our study in §3. We adopt two methods that are originally developed to extract rationales of model predictions (§4.2) for our task and further improve them by adding constraints on the usage of common events and adding prior knowledge of event-level ideology (§4.3).

4.1 Task Overview

We assume our data comes in the form of (a, y), where y is the ideology for article a. We extract events $\mathbf{x} = (x_1, \ldots, x_L)$ from article a where L is the number of events in the article. We define a binary random variable $m_i \in \{0, 1\}$ for each event x_i, and $m_i = 1$ means x_i is a partisan event. The ideology prediction task aims at predicting y using $\mathbf{x}$. The partisan event detection task focuses on predicting partisan indicators $\mathbf{m} = (m_1, \ldots, m_L)$.

4.2 Latent Variable Models

Two-Player Model. We adopt methods in rationale extraction, where rationale is defined as part of inputs that justifies model's prediction (Lei et al., 2016). We specifically use the formulation in Chen et al. (2018), which tackles the rationale (partisan events here) extraction task from an information-theoretic perspective. In details, suppose a positive number k is given, the goal is to extract $k\%$ of events that have the highest mutual information with label y and treat them as partisan events. In other words, our partisan indicator $\mathbf{m}$ satisfies $|\mathbf{m}| = k\% * L$. Since optimizing mutual information is intractable, Chen et al. (2018) provides a variational lower bound as the objective instead:

$$\max_{\mathcal{E}_\theta, q_\phi} \sum_{(\mathbf{x},y)\in\mathcal{D}} \mathbb{E}_{\mathbf{m}\sim\mathcal{E}_\theta(\mathbf{x})} \left[\log q_\phi(y \mid \mathbf{m} \odot \mathbf{x})\right] \tag{2}$$

where $\mathcal{E}_\theta$ is an extractor that models the distribution of $\mathbf{m}$ given $\mathbf{x}$, q_ϕ is a predictor that predicts y given partisan events, $\mathcal{D}$ is the training set, and $\odot$ is the element-wise multiplication.

We parameterize both $\mathcal{E}_\theta$ and q_ϕ using the same model as in §3.2. For the extractor, we first get the embedding $\mathbf{e}$ for all events and then pass it to the transformer encoder to get contextualized event representations. A linear layer converts these representations to logits, from which we sample $k\%$ of them following the subset sampling method in Xie and Ermon (2019)—a differentiable sampling method that allows us to train the whole system end-to-end. At inference time, we select the top $k\%$ of events with the largest logits by the extractor. For the predictor, we again get event embeddings $\mathbf{e}$, but we input $\mathbf{m} \odot \mathbf{e}$ to the transformer encoder so that it only sees the sampled subset of events.

Three-Player Model. Among all events in the article, some may have spurious correlation with the ideology. For instance, the event "a CNN reporter contribute to this article" can almost perfectly reveal article's ideology. To prevent models from focusing on these shortcuts instead of partisan events, we further investigate the method in Yu et al. (2019). Concretely, they propose a three-player model where a third complement predictor q_π^c predicts ideology using the complement of partisan events, i.e., $(\mathbf{1} - \mathbf{m}) \odot \mathbf{x}$. The goal for both predictors is to correctly predict the ideology, i.e., maximize $\log q_\phi(y \mid \mathbf{m}\odot\mathbf{x})$ and $\log q_\pi^c(y \mid (\mathbf{1}-\mathbf{m})\odot\mathbf{x})$. The objective for the extractor is thus to select $k\%$ of events that can predict y while the remaining events cannot as in Eq. 3:

$$\begin{aligned}\max_{\mathcal{E}_\theta} \sum_{(\mathbf{x},y)\in\mathcal{D}} & \mathbb{E}_{\mathbf{m}\sim\mathcal{E}_\theta(\mathbf{x})} \left[\log q_\phi(y \mid \mathbf{m} \odot \mathbf{x})\right. \\ & \left. - \log q_\pi^c(y \mid (\mathbf{1} - \mathbf{m}) \odot \mathbf{x})\right].\end{aligned} \tag{3}$$

Intuitively, the extractor and the complement predictor play an adversarial game, and Eq. 3 drives the extractor to identify partisan events as comprehensive as possible so that the complement predictor cannot perform well. In fact, Yu et al. (2019) uses an explicit objective to penalize $\sum_i m_i$ when it deviates from $k\% * L$, but we find this objective does not work well with Eq. 3, leading to an extractor that either selects all events as partisan events or detects nothing. We thus modify it with the subset sampling method (Xie and Ermon, 2019) again. At inference time, we use q_ϕ for ideology prediction.

Note, both two-player and three-player models can have the single- and multi-article variants, depending on whether the extractor and predictors can access all events in a story or just from a single article. Appendix D details the training process.

4.3 Improving Partisan Event Detection

Restricting Models from Picking Common Events. As shown in Fig. 1, common background events and main events should not be considered as partisan events. We therefore explicitly prohibit models from selecting these events. Precisely, we use the same lexical matching method as in §3.2 to find common events in the story. During training, we add an auxiliary objective that minimizes the probability of the extractor to predict events that appear in both left and right articles as partisan events, thus driving models to prefer events reported by

only one side. We only apply this constraint to multi-article models since it requires story-level context to locate common events.

Pretraining to Add Event Ideology Priori. Prior knowledge, especially the media's stance on controversial topics, plays an important role in partisan content detection. Given that the AllSides training set is relatively small, it is unlikely for the model to gain such knowledge on a broad range of topics. We therefore pretrain a model on BIGNEWSALIGN dataset in Liu et al. (2022) to acquire prior knowledge at the event level.

BIGNEWSALIGN is a dataset with 1 million political news stories, and each story contains about 4 articles that report the same main event. We extract events in these articles and train a DistilRoBERTa model as in §3.2 to predict the ideology of each event, where we use article's ideology as event's ideology. Note that this model takes each event as input and does not consider any context information. Intuitively, it counts the reporting frequency of each event: If an event is reported more by left media, it has a higher probability of being left and vise versa. We use this pretrained model for initialization in the extractor and the predictor.

5 Experiments

Tasks and Datasets. Similar to §3.4, we train all models solely on AllSides and evaluate on AllSides test set, Basil, and our partisan event dataset (PEvent). We measure ideology prediction performance on all three datasets and partisan event detection performance on PEvent.

Evaluation Metrics. For ideology prediction, we measure the macro F1 score at the article level. For partisan event detection, we measure the F1 score for the positive class, i.e., partisan event.

Baselines. We consider the following baselines: (1) We **random**ly predict partisan events with a 0.3 probability, and randomly predict ideology for the article. (2) **Event-prior** is the pretrained event model with ideology priori in §4.3. We run it to get the probability of each event being left and right. We then consider the 30% of events with the most skewed distribution as partisan events. Finally, we take the majority vote among partisan events as article's ideology. Intuitively, this baseline utilizes the prior knowledge of event ideology to detect partisan events. (3) **Non-latent** is the best performing *multi-article* model in §3.4, which does not contain latent variables. Built upon this method, we create

	Ideology Prediction			Event
	AllSides	Basil	PEvent	PEvent
Random	$49.83_{\pm1.65}$	$50.99_{\pm3.40}$	$51.33_{\pm6.79}$	$28.93_{\pm0.23}$
Event-prior	$63.39_{\pm0.00}$	$61.37_{\pm0.00}$	$55.44_{\pm0.00}$	$30.66_{\pm0.00}$
Non-latent-attn	$88.49_{\pm0.74}$	$68.50_{\pm2.07}$	$83.59_{\pm1.67}$	$29.90_{\pm0.63}$
Non-latent-pert	$88.49_{\pm0.74}$	$68.50_{\pm2.07}$	$83.59_{\pm1.67}$	$31.17_{\pm0.99}$
Single-article Models				
Two-player	$66.75_{\pm2.35}$	$59.28_{\pm4.95}$	$48.43_{\pm4.63}$	$28.79_{\pm1.16}$
$+pri.$	$81.50_{\pm0.52}$	$68.65_{\pm2.11}$	$70.87_{\pm2.89}$	$31.53_{\pm0.52}$
Three-player	$66.87_{\pm2.32}$	$60.15_{\pm2.36}$	$48.74_{\pm3.55}$	$29.72_{\pm2.30}$
$+pri.$	$81.06_{\pm0.86}$	$65.60_{\pm0.55}$	$70.57_{\pm2.51}$	$30.70_{\pm2.68}$
Multi-article Models				
Two-player	$86.45_{\pm0.50}$	$69.98_{\pm1.24}$	$82.36_{\pm3.83}$	$33.27_{\pm1.05}$
$+res.$	$85.68_{\pm0.32}$	$68.01_{\pm2.93}$	$82.38_{\pm3.28}$	$\underline{33.54}_{\pm0.91}$
$+pri.$	$\underline{91.03}_{\pm0.72}$	$\underline{71.27}_{\pm1.14}$	$\underline{84.31}_{\pm5.58}$	$33.32_{\pm0.74}$
$+res.+pri.$	$\mathbf{91.58}_{\pm0.25}$	$\mathbf{71.43}_{\pm2.57}$	$\mathbf{89.16}_{\pm3.04}$	$\mathbf{33.99}_{\pm0.39}$
Three-player	$85.48_{\pm1.47}$	$65.08_{\pm1.57}$	$83.10_{\pm1.82}$	$33.20_{\pm3.25}$
$+res.$	$85.84_{\pm0.19}$	$66.81_{\pm2.13}$	$80.36_{\pm3.31}$	$31.68_{\pm1.24}$
$+pri.$	$88.03_{\pm1.19}$	$70.26_{\pm2.94}$	$74.02_{\pm4.03}$	$33.01_{\pm1.62}$
$+res.+pri.$	$88.02_{\pm1.45}$	$69.39_{\pm1.01}$	$80.92_{\pm3.57}$	$31.46_{\pm1.75}$

Table 3: F1 scores (avg. of 5 runs) for ideology prediction and partisan event detection. $res.$: restrict models to prefer events selected by only one side; $pri.$: prior knowledge with pretrained event representations. Models that do cross-article comparison yield better performance on both tasks. Adding prior knowledge helps in almost all cases. Non-latent models have the same ideology prediction scores since they are the same model.

two variants for partisan event detection. The first is **attention**-based method, which is shown effective at finding input words that trigger the sentiment prediction (Wang et al., 2016). We use our trained model and consider the top 30% of events with the largest attention weights (sum over all heads and positions) as partisan events. The second method is **perturbation**-based (Li et al., 2016), where we use the non-latent model and iteratively remove one event at a time and choose 30% of events that lead to the largest output change as partisan events.

Results. Table 3 presents the results when $k = 30$, since 27.28% of events in PEvent dataset are partisan. We explore the influence of k in §6. Unsurprisingly, we observe that multi-article models outperform single-article models on both tasks, emphasizing the importance of story-level context for cross-document event comparison.

On partisan event detection (last column of Table 3), latent variable models outperform all baselines, showing the effectiveness of training with article ideology labels. Moreover, restricting models from selecting common events improves partisan event detection for two-player models, which validates the intuition that common events are less likely to be partisan. Providing prior knowledge of event ideology further boosts on both tasks, especially for single-article models, illustrating the benefits

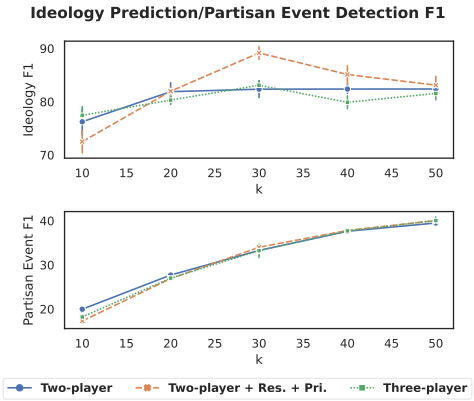

Figure 3: Ideology prediction and partisan event detection performance with different k values (average of 5 runs). Error bars show the standard deviation. Performance variance is small for partisan event detection.

of prior knowledge when the context is limited. Finally, combining the two improvements, the two-player model on the multi-article setup achieves the best performance across the board. It is also important to point out that this model only uses 30% of events to predict ideology, but it still outperforms the model that sees full articles in the story, which suggests that a good modeling of events in the article could be more helpful than raw text representations when predicting ideology.

6 Further Analyses

Effect of Varying k. We now explore the effect of k's values. We experiment with three multi-article models: base two-player, two-player with restriction and prior knowledge, and base three-player models. We train these models with $k = 10, 20, 30, 40, 50$ and plot the performance of partisan event detection and ideology prediction on PEvent in Fig. 3. For event detection, the model improves as k increases, but the improvement is moderate when $k > 30$. For ideology prediction, the performance plateaus at $k = 20$ except for the two-player model with restriction and prior knowledge, which peaks at $k = 30$. This suggests that only a subset of events reflect the article's ideology, and it is enough to make predictions based on them.

Error Analysis. Table 4 and Table 9 in Appendix present the predictions by the two-player model on the multi-article setup with one-sided restriction and prior knowledge for events. Two major types of errors are observed. First, the model struggles when an article attacks a statement from the opposite side with an implicit sentiment. For instance, "threw," "continue," and "had" in Table 4 are events or statements from the right, but the author reports them with an implicit negative sentiment (e.g., "not a thing!"), making the event flatter to the left. Future models need to have an enhanced understanding of implicit sentiment along with the involving entities (Deng and Wiebe, 2015; Zhang et al., 2022). Second, the model still frequently selects main events as partisan content, as shown by the "delivered" event in Table 9, maybe because models need to include it as necessary context. Future work may consider adding cross-document event coreference (Cybulska and Vossen, 2014).

Title: At the NRA Convention, People **Blame** Mass Shootings on Everything But Guns

The nation has been plunged into despair and mourning . . . in Houston, the National Rifle Association still **threw** a party . . . Two messages **emerged** from the assembled throngs and the doting politicians in attendance, just 300 miles from Uvalde: 1) People must **continue** to enjoy the right to acquire any damn firearm they choose, without meddling from the state; and 2) the massacre **had** absolutely nothing–not a thing!–to do with the untrammeled commerce in guns . . .

Ideology label: left **Prediction**: left

Table 4: Article snippets of model predictions (multi-article two-player model with both improvements) and annotations. Colored spans denote events, with the **predicate** bolded. Blue: model predictions; red: human annotations; purple: annotations and predictions.

7 Conclusion

Partisan event selection is an important form of media bias which may exist in even the most apparently nonpartisan news, but which is especially hard to detect without extensive cross-article comparison. We first verify the existence of partisan event selection by inspecting the impact of partisan events on the performance of ideology prediction. We then jointly detect partisan events and predict an article's ideology using latent variable models. Experiments show that our models identify partisan events that reasonably align with human judgement, and our model using cross-article event context outperforms a baseline that only considers events in single articles, on both partisan event detection and ideology prediction. In addition to detecting partisan events and underlying ideologies, our analysis suggests future directions for identifying interactions among entities in an article and for event coreference across articles, to provide more complete and ideologically varied news stories.

8 Limitations

We investigate the impact of event selection on models' ideology prediction performance, to verify the existence of event selection in news media. The results, however, do not state a causal relation between media ideology and reported events.

We analyze the model output and discuss in details two major limitations of our latent variable models in §6. Apart from those two errors, we also observe that events detected by the model as partisan may not align with the model's prediction of the article's ideology. In other words, the model could identify right-leaning events as partisan events while predicting the article as left-leaning (Table 5). Although the methods we adopt in this paper identify events that are indicative of ideology (Chen et al., 2018; Yu et al., 2019), they do not provide further justifications for how these events interact to reflect the ideology. For instance, the extractor could detect a right event and several left events that attack it. To further understand the event selection effect, future work may consider incorporating event-level ideology to model the interplay among events.

Although our models that include cross-article context can be extended to any number of articles without modification, they may be restricted by the GPU memory limit in practice. Particularly, the Transformer encoder that contextualizes all events in a story requires computational resources to scale quadratically with the number of events, which is infeasible for stories that contain many articles. Future work may consider designing special attention patterns based on the discourse role of each event in the article (van Dijk, 1988; Choubey et al., 2020).

Finally, due to the cost of manual labeling, we only evaluate our partisan event detection models on a dataset that covers two specific political issues. It remains to be seen whether methods introduced in this paper can be generalized to a broader range of issues. We call for the community's attention to design and evaluate partisan event detection models on more diverse topics.

9 Ethical Considerations

9.1 Dataset Collection and Usage

Partisan Event Dataset Collection. We conform with the terms of use of the source websites and the intellectual property and privacy rights of the original authors of the texts when collecting articles.

Title: Biden calls for assault weapons ban, making gun manufacturers liable for shootings President Biden on Thursday made an emotional appeal for ambitious new gun laws, including a ban on military-style rifles . . . On the other side of the aisle, Republicans **bristled** at Democrats' equating support for the Second Amendment with tolerating mass murder. "You think we don't have hearts," said Rep. Louis Gohmert, Texas Republican.
Ideology label: right **Prediction**: left

Table 5: Article snippets where the extractor detects a right event, but the predictor predicts the article as left.

We do not collect any sensitive information that can reveal original author's identity. We also consult Section 107[6] of the U.S. Copyright Act and ensure that our collection action fall under the fair use category.

Datasets Usage. Except the partisan event dataset collected in this work, we get access to the Basil dataset by direct download. For AllSides, we contact with the authors and obtain the data by agreeing that we will not further distribute it.

9.2 Usage in Application

Intended Use. The model developed in this work has the potential to assist the public to better understand and detect media bias in news articles. The experiments in §5 show that our model is able to identify partisan events on two controversial issues that moderately align with human judgement. The detected events can be presented to show different perspectives from both ends of the political spectrum, thus providing readers with a more complete view of political issues.

Failure Modes. Our model fails when it mistakenly predicts a non-partisan event as a partisan event, misses out the partisan events, or predicts the wrong ideology for an article. They may cause misperception and misunderstanding of an event. For vulnerable populations (e.g., people who maybe not have the specific knowledge to make the right judgements), the harm could be amplified if they blindly trust the machine outputs.

Biases. The training dataset is roughly balanced in the number of left and right articles, so the model is not trained to encode bias. However, the dataset is relatively small and does not cover all possible

[6] https://www.copyright.gov/title17/92chap1.html#107

political topics. Particularly, most of the news articles in the training set are related to U.S. politics, thus the model is not directly applicable to other areas in the world.

Misuse Potential. Users may mistakenly take the model outputs as ground truth. We recommend any usage of our model displaying an "use with caution" message to encourage users to cross-check the information from different sources and not blindly trust a single source.

References

Amr Ahmed and Eric Xing. 2010. Staying informed: Supervised and semi-supervised multi-view topical analysis of ideological perspective. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1140–1150, Cambridge, MA. 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.

Matthew A. Baum and Tim Groeling. 2008. New media and the polarization of american political discourse. *Political Communication*, 25(4):345–365.

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

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

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

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 438–444, Beijing, China. Association for Computational Linguistics.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*, pages 789–797, Columbus, Ohio. Association for Computational Linguistics.

Jianbo Chen, Le Song, Martin J. Wainwright, and Michael I. Jordan. 2018. Learning to explain: An information-theoretic perspective on model interpretation. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 882–891. PMLR.

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 Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Agata Cybulska and Piek Vossen. 2014. Using a sledgehammer to crack a nut? lexical diversity and event coreference resolution. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4545–4552, Reykjavik, Iceland. European Language Resources Association (ELRA).

Dave D'Alessio and Mike Allen. 2006. Media Bias in Presidential Elections: A Meta-Analysis. *Journal of Communication*, 50(4):133–156.

Claes de Vreese. 2004. The effects of strategic news on political cynicism, issue evaluations, and policy support: A two-wave experiment. *Mass Communication and Society*, 7(2):191–214.

Stefano DellaVigna and Matthew Gentzkow. 2009. Persuasion: Empirical evidence. Working Paper 15298, National Bureau of Economic Research.

Stefano DellaVigna and Ethan Kaplan. 2007. The Fox News Effect: Media Bias and Voting*. *The Quarterly Journal of Economics*, 122(3):1187–1234.

Lingjia Deng and Janyce Wiebe. 2015. MPQA 3.0: An entity/event-level sentiment corpus. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1323–1328, Denver, Colorado. Association for Computational Linguistics.

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

Robert M. Entman. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4):51–58.

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.

Anjalie Field, Doron Kliger, Shuly Wintner, Jennifer Pan, Dan Jurafsky, and Yulia Tsvetkov. 2018. Framing and agenda-setting in Russian news: a computational analysis of intricate political strategies. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3570–3580, Brussels, Belgium. Association for Computational Linguistics.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia. Association for Computational Linguistics.

Matthew Gentzkow and Jesse M. Shapiro. 2006. Media bias and reputation. *The 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, Jesse M Shapiro, and Michael Sinkinson. 2014. Competition and ideological diversity: Historical evidence from us newspapers. *American Economic Review*, 104(10):3073–3114.

Matthew Gentzkow, Jesse M. Shapiro, and Daniel F. Stone. 2015. Chapter 14 - media bias in the marketplace: Theory. In Simon P. Anderson, Joel Waldfogel, and David Strömberg, editors, *Handbook of Media Economics*, volume 1 of *Handbook of Media Economics*, pages 623–645. North-Holland.

Sean M. Gerrish and David M. Blei. 2011. Predicting legislative roll calls from text. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, page 489–496, Madison, WI, USA. Omnipress.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Justin Grimmer. 2010. A bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. *Political Analysis*, 18(1):1–35.

Guy Grossman, Yotam Margalit, and Tamar Mitts. 2022. How the ultrarich use media ownership as a political investment. *The Journal of Politics*, 84(4):1913–1931.

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.

Yeooul Kim, Suin Kim, Alejandro Jaimes, and Alice Oh. 2014. A computational analysis of agenda setting. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14 Companion, page 323–324, New York, NY, USA. Association for Computing Machinery.

Vivek Kulkarni, Junting Ye, Steve Skiena, and William Yang Wang. 2018. Multi-view models for political ideology detection of news articles. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3518–3527, Brussels, Belgium. Association for Computational Linguistics.

Valentino Larcinese, Riccardo Puglisi, and James M Snyder Jr. 2011. Partisan bias in economic news: Evidence on the agenda-setting behavior of us newspapers. *Journal of public Economics*, 95(9-10):1178–1189.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Austin, Texas. Association for Computational Linguistics.

Chang Li and Dan Goldwasser. 2019. Encoding social information with graph convolutional networks forPolitical perspective detection in news media. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2594–2604, Florence, Italy. Association for Computational Linguistics.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. Understanding neural networks through representation erasure.

Siyi Liu, Lei Guo, Kate Mays, Margrit Betke, and Derry Tanti Wijaya. 2019a. Detecting frames in news headlines and its application to analyzing news framing trends surrounding U.S. gun violence. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 504–514, Hong Kong, China. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach.

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. In *International Conference on Learning Representations*.

Maxwell McCombs and Amy Reynolds. 2008. How the news shapes our civic agenda. In *Media Effects*, pages 17–32. Routledge.

Maxwell E. McCombs and Donald L. Shaw. 1972. The agenda-setting function of mass media. *The Public Opinion Quarterly*, 36(2):176–187.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

Sendhil Mullainathan and Andrei Shleifer. 2005. The market for news. *American Economic Review*, 95(4):1031–1053.

Viet-An Nguyen, Jordan L Ying, and Philip Resnik. 2013. Lexical and hierarchical topic regression. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Qiang Ning, Sanjay Subramanian, and Dan Roth. 2019. An improved neural baseline for temporal relation extraction. 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 6203–6209, Hong Kong, China. Association for Computational Linguistics.

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.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. Curran Associates Inc., Red Hook, NY, USA.

Elizabeth M. Perse and Jennifer Lambe. 2016. *Media Effects and Society*. Routledge.

Gerald Prince. 2012. A grammar of stories. In *A Grammar of Stories*. De Gruyter Mouton.

James Pustejovsky, José M. Castaño, Robert Ingria, Roser Saurí, Robert J. Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R. Radev. 2003. Timeml: Robust specification of event and temporal expressions in text. In *New Directions in Question Answering*.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Kevin M. Quinn, Burt L. Monroe, Michael Colaresi, Michael H. Crespin, and Dragomir R. Radev. 2010. How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, 54(1):209–228.

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

Peng Shi and Jimmy Lin. 2019. Simple bert models for relation extraction and semantic role labeling. *ArXiv*, abs/1904.05255.

Oren Tsur, Dan Calacci, and David Lazer. 2015. A frame of mind: Using statistical models for detection of framing and agenda setting campaigns. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1629–1638, Beijing, China. Association for Computational Linguistics.

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 Adrianus van Dijk. 1988. *News as discourse*. L. Erlbaum Associates.

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

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based LSTM for aspect-level sentiment classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 606–615, Austin, Texas. Association for Computational Linguistics.

Sean Wilner, Daniel Woolridge, and Madeleine Glick. 2021. Narrative embedding: Re-Contextualization through attention. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1393–1405, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Winston Wu and David Yarowsky. 2020. Computational etymology and word emergence. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3252–3259, Marseille, France. European Language Resources Association.

Sang Michael Xie and Stefano Ermon. 2019. Reparameterizable subset sampling via continuous relaxations. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, IJCAI'19, page 3919–3925. AAAI Press.

Mo Yu, Shiyu Chang, Yang Zhang, and Tommi Jaakkola. 2019. Rethinking cooperative rationalization: Introspective extraction and complement control. 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 4094–4103, Hong Kong, China. Association for Computational Linguistics.

Shuaicheng Zhang, Lifu Huang, and Qiang Ning. 2021. Extracting temporal event relation with syntactic-guided temporal graph transformer.

Xinliang Frederick Zhang, Nick Beauchamp, and Lu Wang. 2022. Generative entity-to-entity stance detection with knowledge graph augmentation. *CoRR*, abs/2211.01467.

Hyperparameter	Value
number of epochs	20
patience	4
maximum learning rate	3e-5
learning rate scheduler	linear decay with warmup
warmup percentage	6%
optimizer	AdamW (Loshchilov and Hutter, 2019)
weight decay	5e-5
# FFNN layer	2
hidden layer dimension in FFNN	768
dropout in FFNN	0.1

Table 6: Hyperparameters used for the event extraction model.

Appendix A Implementation Details

For all experiments in this paper, our implementation is based on Pytorch (Paszke et al., 2019) and HuggingFace transformers (Wolf et al., 2020) library, and we preprocess all articles using Stanza (Qi et al., 2020). All experiments are conducted on 4 NVIDIA RTX A6000 GPUs.

Appendix B Event-based Ideology Prediction Models

B.1 Event Extraction

We train a RoBERTa-large model (Liu et al., 2019b) on MATRES dataset (Ning et al., 2018) to predict whether each token is an event predicate or not. To provide surrounding context, we split articles into groups of 4 sentences and process 4 sentences together. We follow previous work on using TimeBank and AQUAINT sections in MATRES as training set and Platinum section as test set (Ning et al., 2019). Table 6 shows the hyperparameters for model architecture and training process.

B.2 Contextualized Event Representation

Here we detail our model that uses cross-article context for ideology prediction. As described in §3.2, we first input the sentence that contains the event to a DistilRoBERTa model (Sanh et al., 2019) to get event representation $\mathbf{e}$. This representation is then passed to a Transformer encoder (Vaswani et al., 2017) with three embeddings to obtain contexualied event representation $\mathbf{c}$:

- **Article embedding** indicates the index of the article that contains the event, with one embedding per article index. The datasets we experiment with in this paper have at most 3 articles in each story. During training, we randomly shuffle the articles in each story.
- **Frequency embedding** informs the model whether the event appears in only one article, at least two but not all articles, or all articles in the story. We have one embedding per category. We find common events through lexical matching. Concretely, we use a dictionary that contains derivational morphology mappings (Wu and Yarowsky, 2020) to get the base form of the event predicate. We then construct a set of words for the predicate by including the synonyms for the base form and original form (Bird et al., 2009). Finally, two events are considered as the same if their predicate sets overlap and both of their ARG0 and ARG1 have a high word overlap (a threshold of 0.4,[7] calculated by overlap coefficient, without stop words).
- **Position embedding** represents the relative position of the event in the article. We multiply the relative position of the event (a real number in $[0, 1]$) with a learnable embedding.

We further train a `[SEP]` token that separates the events from different articles. Finally, average representation of all events in an article is used to predict the article's ideology. Table 7 includes the hyperparameters of the model.

The entire model contains 106M parameters. On average, the training takes 25 minutes on a single NVIDIA RTX A6000 GPU.

Appendix C Partisan Event Annotation

Data Collection. We manually collect 25 stories, each with three articles from AllSides[8] that relate to the mass shooting in Texas and the overturn of *Roe v. Wade*. We extract events from each article and only keep the left and right article in each story.[9] We mask out the name of the media (e.g., "CNN" and "Fox News") in the article before annotation to avoid bias.

[7] We search threshold values from 0.2 to 0.5 by manually inspecting identified common events in 6 articles. A value of 0.4 can identify common events accurately while still allowing variations such as variants of mentions (e.g., president vs. president Biden).

[8] https://www.allsides.com/unbiased-balanced-news

[9] Each story on AllSides contains three articles from left, center, and right respectively. We only include the left and right articles in our dataset.

Hyperparameter	Value
number of epochs	5
maximum learning rate	5e-5
learning rate scheduler	linear decay with warmup
warmup percentage	6%
optimizer	AdamW
weight decay	1e-4
transformer hidden dimension	768
transformer # heads	12
# transformer layer	4
# FFNN layer	2
hidden layer dimension in FFNN	768
dropout in FFNN	0.1

Table 7: Hyperparameters used for the event-based ideology prediction model.

Annotation Process. We hire three college students proficient in English and familiar with discerning ideology under the context of U.S. political spectrum. We present each story, together with extracted events (predicate, ARG0, and ARG1) to annotators, without revealing the media source. The annotators are asked to first finish reading two articles on the same story but written by media of left and right leanings. They will then answer the following questions:

- Sort articles by their ideological position (left or right) in this story.
- Based on the ideological leaning that you labeled in the above question, select the **partisan events** that you think reflect the ideology of the article. Please note: (1) If you label an article as left, only select partisan events that favor the left side. (2) If an event appears in both articles as the main events or as the necessary background events, and it does not contain salient sentiment, do not select it as a partisan event.

Two annotators label all 50 articles, and a third annotator compares their annotations and resolves conflicts. We calculate inter-annotator agreement on all 50 articles and numbers can be found in §3.3.

Partisan Events Distribution. We further investigate the distribution of position of partisan events in the article. Fig. 4 shows the percentage of partisan events that belong to each quartile of an article. As can be observed, right articles have more partisan events that appear in later parts of an article, whereas partisan events in left articles are evenly distributed in the article.

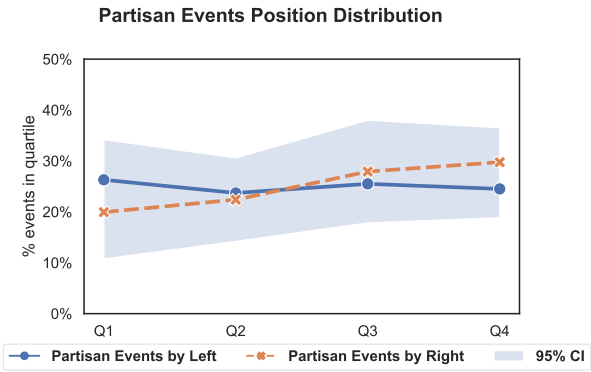

Figure 4: Distribution of partisan events found in each quartile of an article, in terms of spatiality. Shaded area shows the 95% confidence interval.

	# articles	# events
Left	128,481	6,280,732
Right	123,380	4,986,165

Table 8: Statistics for the BIGNEWSALIGN pretraining dataset.

Appendix D Latent Variable Models

Implementation Details. For both extractor and predictors, we use the same model architecture as in §B.2 with hyperparameters listed in Table 7. For the three-player model, we follow the training process in Generative Adversarial Nets training (Goodfellow et al., 2014).

The two-player model contains 213M parameters, and the three-player model contains 320M parameters. On average, the training takes 50 minutes for the two-player model and 1.5 hours for the three-player model on a single NVIDIA RTX A6000 GPU.

Pretrained Model for Event Representation. We use the BIGNEWSALIGN dataset (Liu et al., 2022) to pretrain a model with prior event ideology knowledge. We remove stories in the dataset that contain duplicate articles and downsample articles in each story so that the number of left and right articles are balanced. Table 8 shows the statistics of the pretraining dataset. We then train a DistrilRoBERTa model that takes each event as input and predicts the event's ideology, where we use the article ideology as the event's ideology. We train this model on BIGNEWSALIGN for 2 epochs and use it to initialize our latent variable models.

Title: "Enough": Biden **Exhorts** Congress To **Pass** Gun Control Laws President Joe Biden **delivered** the second evening address of his presidency on Thursday night, almost **begging** Congress to **pass** gun control legislation . . . However, Biden **cited** former Supreme Court Justice Antonin Scalia-a conservative icon-who had declared that the Second Amendment was "not unlimited."
Ideology label: left **Prediction**: left

Table 9: Article snippets of human annotations and model predictions (multi-article two-player model with both improvements). Highlighted spans denote events, with the **predicate** bolded. Blue: model predictions; purple: human annotations and predictions.

Appendix E Additional Error Analysis

Table 9 in this section is supplementary to the Error Analysis section in §6. The model detects "delivered" as a partisan event, which is the main event in the story. The constraint introduced in §4.3 fails in this case because the other article describes this event differently (i.e., "Biden made an emotional appeal"), thus suggesting future research direction that leverages cross-document event coreference.