Using Social and Linguistic Information to Adapt Pretrained Representations for Political Perspective Identification

Chang Li
Department of Computer Science
Purdue University, West Lafayette, IN
li1873@purdue.edu

Dan Goldwasser
Department of Computer Science
Purdue University, West Lafayette, IN
dgoldwas@purdue.edu

Abstract

Understanding the political perspective shaping the way events are discussed in the media is increasingly important due to the dramatic change in news distribution. With the advance in text classification models, the performance of political perspective detection is also improving rapidly. However, current deep learning based text models often require a large amount of supervised data for training, which can be very expensive to obtain for this task. Meanwhile, models pre-trained on the general source and task (e.g. BERT) lack the ability to focus on bias-related text span. In this paper, we propose a novel framework that pre-trains the text model using signals from the rich social and linguistic context that is readily available, including entity mentions, news sharing, and frame indicators. The pre-trained models benefit from tasks related to bias detection and therefore are easier to train with the bias labels. We demonstrate the effectiveness of our proposed framework by experiments on two news bias datasets. The models with pre-training achieve significant improvement in performance and are capable of identifying the text span for bias better.

1 Introduction

The perspectives underlying the way information is conveyed to readers can prime them to take similar stances and shape their worldview (Gentzkow and Shapiro, 2010, 2011). Given the highly polarized coverage of news events, recognizing these perspectives can help ensure that all points of view are represented by news aggregation services, and help avoid "information echo-chambers" in which only a single viewpoint is represented. It may also help to prevent the spread of false information online by showing people news with different perspectives.

Past work studying the expression of bias in the text has focused on lexical and syntactic representations of bias (Greene and Resnik, 2009; Recasens et al., 2013; Elfardy et al., 2015). Expressions of bias can include the use of the passive voice (e.g., *"mistakes were made"*), or references to known ideological talking points (Baumer et al., 2015; Budak et al., 2016; Card et al., 2016; Field et al., 2018; Morstatter et al., 2018) (e.g., *"pro-life"* vs. *"pro-choice"*). However, bias in news media is often nuanced and very difficult to detect. Journalists often strive to appear impartial and use language that does not reveal their opinions directly. Also, by their nature, news articles describing the same real-world event will share many similar details of the event, regardless of their political perspectives. Instead, bias is often expressed through informational choices (Fan et al., 2019), which highlight different aspects of the news story and frame facts shared by all articles in different ways. For example, the following articles capture different perspectives (Top *left*, Bottom *right*), while discussing the same news event– the 2021 storming of the U.S. Capitol [1].

Adapted from **NYTimes** (*Left*)

How Republicans Are Warping Reality Around the Capitol Attack ... Jim Hoft, did not reply to questions but did send along several of his own news articles related to claims of antifa involvement in the Capitol attack — citing the case of a man named **John Sullivan**, whom the right-wing media has dubbed an **"antifa leader"** in efforts to prove its theory of infiltration.

Adapted from **Fox News** (*Right*)

BLM activist inside Capitol claims he was 'documenting' riots, once said 'burn it all down'. John Sullivan has previously called for 'revolution' and to 'rip Trump' out of his office. An anti-Trump activist who once said he wanted to "rip" the president out of office entered the Capitol Building Wednesday alongside a mob of pro-Trump protesters, but he said he was just there to "document" it.

The two articles discuss the presentation of *John*

[1] https://en.wikipedia.org/wiki/2021_storming_of_the_United_States_Capitol

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 4569–4579

Sullivan as an Antifa member[2] who participated in the Capitol storming. However the story is framed in very different ways - while the bottom article frames the story directly as a discussion of Antifa involvement, the top discusses it in the context of political messaging and journalism. Furthermore, we notice that the difference is focused on a specific entity - John Sullivan.

Despite the fact that these distinctions are easily detectable by a human reader familiar with the political divisions in the U.S., they are very difficult to detect automatically. Recent success stories using large-scale pre-training for constructing highly expressive language models (Devlin et al., 2019) are designed to capture co-occurrence patterns, likely to miss these subtle differences.

In this paper we suggest that bias detection requires a different set of self-supervised pre-training objectives that can help provide a better starting point for training downstream biased detection tasks. Specifically, we design three learning objectives. The first, captures *political knowledge*, focusing on the embedding of political entities discussed in the text. The second one captures *external social context*. Following the intuition that different social groups would engage with documents expressing a different bias (e.g., left-leaning users are more likely to read the NYTimes article compared to the Fox News article), we collect social information contextualizing news articles and learn to predict the social context of each article, based on its content, thus aligning the two representations. Finally, the third is based on linguistic knowledge, focusing on the *issue framing* decisions made by the authors. Framing decisions have been repeatedly shown to capture political bias (Recasens et al., 2013; Johnson and Goldwasser, 2016; Roy and Goldwasser, 2020; Mendelsohn et al., 2021), and we argue that infusing a language model with this information can help capture relevant information. Note that this information is only used for pre-training. Other works using social information to analyze political bias (Li and Goldwasser, 2019; Nguyen et al., 2020; Pacheco and Goldwasser, 2021) augment the text with social information, however since this information can be difficult to obtain in real-time, we decided to investigate if it can be used as a distant supervision source for pre-training a language model.

These pre-training tasks are then used for training a **M**ulti-head **A**ttention **N**etwork (MAN) which creates a bias-aware representation of the text.

We conducted our experiments over two datasets, Allsides (Li and Goldwasser, 2019) and SemEval Hyperpartisan news detection (Kiesel et al., 2019). We compared our approach to several competitive text classification models and conducted a careful ablation study designed to evaluate the individual contribution of pre-training through knowledge from various contexts. Our results demonstrate the importance of all aspects, each contributing to the model's performance.

2 Related Work

The problem of perspective identification is originally studied as a text classification task (Lin et al., 2006; Greene and Resnik, 2009; Iyyer et al., 2014), in which a classifier is trained to differentiate between specific perspectives. Other works use linguistic indicators of bias and expressions of implicit sentiment (Recasens et al., 2013; Baumer et al., 2015; Field et al., 2018).

Recent work by (Fan et al., 2019) aims to characterize content relevant for bias detection. Unlike their work which relies on annotated spans of text, we aim to characterize this content without explicit supervision.

In the recent SemEval-2019, a hyperpartisan news article detection task was suggested[3]. Many works attempt to solve this problem with deep learning models (Jiang et al., 2019; Hanawa et al., 2019). We build on these works to help shape our text representation approach.

Several recent works also started to make use of concepts or entities appearing in the text to get a better representation. Wang et al. (2017) treats the extracted concepts as pseudo words and appends them to the original word sequence which is then fed to a CNN. The KCNN model by Wang et al. (2018), used for news recommendation, concatenates entity embeddings with the respective word embeddings at each word position to enhance the input. We take a different approach and instead try to inject knowledge of entities into the text model through the masked entity training. Zhang et al. (2019) also uses entity-level masking for training. However, they predict the tokens for the masked

[2] https://en.wikipedia.org/wiki/Antifa_(United_States)

[3] https://pan.webis.de/semeval19/semeval19-web/

entity instead of relying on meaningful representations for entities as ours.

Political framing, due to its relation with ideology and perspective, is studied in the NLP communities (Johnson et al., 2017; Field et al., 2018; Shurafa et al., 2020). There is also growing interest in utilizing framing differences to identify bias in news articles (Roy and Goldwasser, 2020).

Pre-trained models are widely used in numerous NLP tasks, from the early word2vec representation (Mikolov et al., 2013) to the generic language models like ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019). Recently, people also started to work on task specific pre-training that try to bring task and domain related knowledge into the model. Xu et al. (2019) is similar to our work as it proposes to enhance the BERT model through training on review data and sentiment classification tasks so that it can obtain better performance across multiple review-based tasks.

3 Political Perspective Identification Task

The problem of political perspective identification in news media can be formalised as follows. Given a news article d, where d consists of sentences s_i, $i \in [1, L]$, and each sentence s_i consists of words w_{it}, $t \in [1, T]$. L and T are the number of sentences in d and number of words in s_i respectively. The goal of this task is to predict the political perspective y of the document. Given different datasets, this can either be a binary classification task, where $y \in \{0, 1\}$ (hyperpartisan or not), or a multi-class classification problem, where $y \in \{0, 1, 2\}$ (left, center, right).

The overall architecture of our model is shown in Figure 1. It includes two sequence encoders, one for word level and another for sentence level. The hidden states from an encoder are combined through a multi-head self-attention mechanism. With pre-training on various social and linguistic information, the generated sentence and document vectors will consider not only the context within the text but also the knowledge about the entities (e.g. their political affiliation, or stance on controversial issues), sharing users, and frame indicators. We explain the structure of our model and the rich social and linguistic context we consider in detail below. Note that our pre-training strategies are not tied with any specific model structure and can be easily applied to other text models.

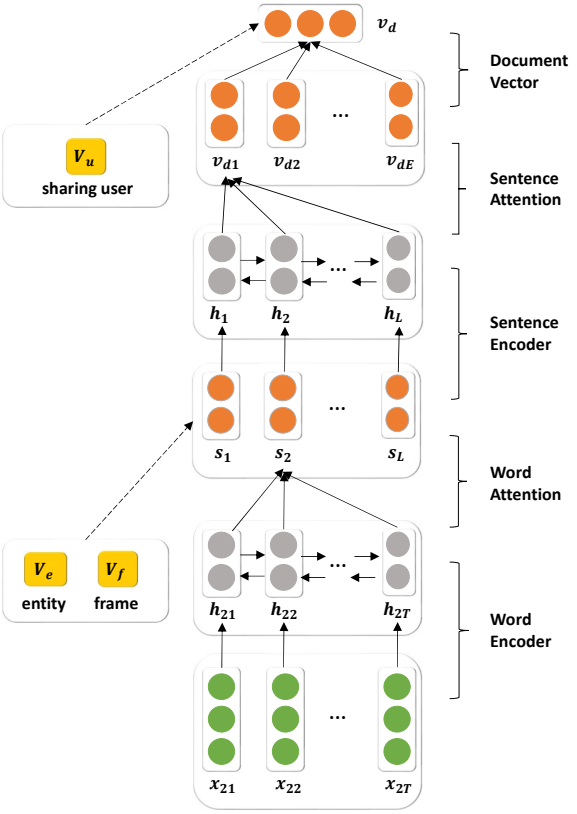

Figure 1: Overall Architecture of MAN Model.

3.1 Multi-Head Attention Network

The basic component of our model is the Hierarchical LSTM model (Yang et al., 2016). The goal of our model is to learn document representation v_d for political perspective prediction. It consists of several parts: a word sequence encoder, a word-level attention layer, a sentence sequence encoder, and a sentence-level attention layer. To capture the context in both directions, we use bidirectional LSTM in this work. For each element in the input sequence, the hidden state h is a concatenation of the forward hidden state $\overrightarrow{h}$ and backward hidden state $\overleftarrow{h}$ computed by the respective LSTM cells.

Given a sentence with words w_{it}, $t \in [1, T]$, each word is first converted to its embedding vector x_{it}. We can adopt pre-trained Glove (Pennington et al., 2014) word embeddings or deep contextualized word representation ELMo (Gardner et al., 2017) for this step. The word vectors are then fed into a word level bidirectional LSTM network to incorporate contextual information within the sentence. The hidden states h_{it} from the bidirectional LSTM network are passed to the next layer. In (Yang et al., 2016), a self attention mechanism is introduced to identify words that are important to

the meaning of the sentence, and therefore higher weights are given to them when forming the aggregated sentence vector.

$$p_{itw} = tanh(W_w h_{it} + b_w) \quad (1)$$

$$\alpha_{itw} = \frac{exp(p_{itw}^T p_w)}{\sum_t exp(p_{itw}^T p_w)} \quad (2)$$

$$s_{iw} = \sum_t \alpha_{itw} h_{it} \quad (3)$$

p_{itw} encode the importance of a specific word according to its context, which is compared with the word level preference vector p_w to compute a similarity score. The scores are then normalized to get the attention weight α_{itw} through a softmax function. A weighted sum of the word hidden states is computed based on the attention weight as the sentence vector s_{iw}.

Inspired by the multi-head attention scheme in (Vaswani et al., 2017), we propose a multi-head attention in our model to extend its ability to jointly attend to information at different positions. The sentence vector s_i is computed as an average of s_{iw} obtained from different attention heads. Note that we learn a separate copy of the parameters W_w, b_w and p_w for each attention head.

$$s_i = \frac{\sum_w s_{iw}}{NH_W} \quad (4)$$

where NH_W is the number of word-level attention heads.

Given the sentence vectors s_i, we can generate the document vector v_d in a similar way. It captures the bias related information in news articles and can be used as features for predicting the document bias label.

$$f_d = W_c v_d + b_c \quad (5)$$

$$p_d = \text{softmax}(f_d) \quad (6)$$

We use the negative log likelihood of the correct labels as classification training loss:

$$L = -\sum_d \log p_{dj} \quad (7)$$

where j is the bias label of d.

3.2 Political Entities

News articles, especially the ones we are interested in in this work, are mainly covering real-world events involving political entities and their relations. To better understand the stance over controversial issues and the underlying ideology reflected in the text, it is very important to have extensive world knowledge about these entities, including their traits, opinions, and relevant events. We obtain the entity knowledge representations through learning on Wikipedia data.

Wikipedia2Vec (Yamada et al., 2018) is a model that learns entity embeddings from Wikipedia. It learns embeddings of words and entities by iterating over the entire Wikipedia pages and maps similar words and entities close to one another in a continuous vector space. It jointly optimizes the following three submodels:

1. Wikipedia link graph model, which learns entity embeddings by predicting neighboring entities in Wikipedia's link graph, an undirected graph whose nodes are entities and edges represent links between entities in their Wikipedia pages.

2. Word-based skip-gram model, which learns word embeddings by predicting neighboring words given each word on a Wikipedia page.

3. Anchor context model, which aims to place similar words and entities near one another in the vector space. The objective here is to predict neighboring words given each entity referred to on a Wikipedia page.

The learned entity embeddings encode the background knowledge about these entities in Wikipedia, such as gender, ideology, among others. We use them to initialize our entity embeddings in Section 4.1 which enables us to inject background knowledge of entities to the text model through pre-training.

3.3 Social Information Graph

With the great popularity of social media platforms, many people nowadays tend to share their personal interests and opinions and exchange ideas about social events with others online. This also applies to the sharing of news articles on social media. Intuitively, news articles shared by the same user are likely to have the same bias, and users who share a lot of news in common are close in their

political preferences as well. Hence, we can use this information to guide the pre-training of our text model.

We follow the work in (Li and Goldwasser, 2019) to learn the embeddings through the structure of the social information graph for users who share articles. The graph consists of three types of vertices, namely political users, sharing users, and news articles. Political users are famous politicians or journalists with a clear, self-reported political bias. Sharing users are Twitter users who shared news articles in the dataset. There are two types of edges: 1) following edge between a sharing user to a political user and 2) sharing edge between a sharing user to a news article). Graph Convolutional Networks (GCN) are used to model the graph structure to predict the bias of political users. It aggregates information from the local neighborhood for each node in the graph. Therefore the training of GCN helps to propagate political preference information from political users to sharing users. We use the learned embeddings to guide the pre-training in Section 4.2 so that our text model can use this as distant supervision to map the representation of news articles shared by the same user to be close in the vector space since they are more likely to have the same perspective.

3.4 Frame Indicators

Political framing, studied by political scientists, provides a useful way to study different political perspectives. The frames surrounding an issue can change the reader's perception without having to alter the actual facts as the same information is used as a base. It is a political strategy that used to bias the discussion on an issue toward a specific stance. For example, regarding the topic of abortion, the liberal side will highlight the freedom of choice for women to decide whether to terminate a pregnancy while the conservative side may emphasize the morality aspect instead, arguing the right of the fetus.

Previous work (Roy and Goldwasser, 2020) shows that frame indicators can be used to identify the political perspectives effectively for different topics. These are words that have high pointwise mutual information with a specific frame. They can be considered to represent a more detailed point within a frame. Therefore we propose to use these frame indicators to guide the pre-training of text models so that they can learn to distinguish the nuance between different frames and talking points.

4 Pre-training

As discussed in the introduction, the supervision on news bias requires a lot of human effort to get. Moreover, the text model trained only on the political perspective labels cannot benefit from the rich knowledge we have from the various social and linguistic contexts presented in the previous section. To enhance the performance of political perspective identification, we may need to bring external knowledge and signals from the aforementioned contexts to enable the text model to take them into account when processing the news article. Eventually, we want to show that the model works best by exploiting all different kinds of knowledge and signals related to the task.

4.1 Entity Guided Pre-training

The goal of entity-guided pre-training is to inject knowledge about entities into our text model to help solve the political perspective identification problem. We first extract entities from the data corpus and then learn knowledge representations for them using Wikipedia2Vec introduced in 3.2. We then use the predicted

We utilize the entity linking system DBpedia Spotlight (Daiber et al., 2013) to recognize and disambiguate the entities in news articles. We use the default configuration of DBpedia Spotlight, including the confidence threshold of 0.35, which helps to exclude uncertain or wrong entity annotations. We keep only entities with Person or Organization types that appear in the corpus.

Inspired by the masked language modeling objective used in BERT (Devlin et al., 2019), we propose an entity-level masking task for injecting background knowledge of entities into the text model based on the news articles in which they are mentioned. The objective is to predict the masked entity based on the context provided by the other words in a sentence. Specifically, the entity mentions (regardless of the number of tokens in text) are replaced with a special token "[MASK]" during preprocessing. We use a bidirectional LSTM (sentence level encoder described in 3.1) to encode the sentence, and the hidden state of the mask token will be used for prediction. We use negative sampling to randomly generate negative entity candidates from all entities in our dictionary uniformly. The prediction can be done by comparing

the similarity score between the hidden state and the embedding of candidate entities mapped to the same space through a hidden layer.

$$h_{it}^T \cdot (W_e v_e + b_e) \tag{8}$$

where h_{it} is the hidden state for the masked token, v_e the embedding of entity e, W_e and b_e the parameters for the mapping hidden layer. We use the multi-class cross-entropy loss for all pre-training tasks.

The learned sentence encoder will then be able to highlight the context in the news articles that is more related to the properties and of the mentioned entities.

4.2 Sharing User Guided Pre-training

As we discussed in Section 3.3, the sharing behavior by Twitter users can be regarded as signals to guide the pre-training of our text model. In order to benefit from the social information available, we propose to predict the sharing user given a news article. Similar to the previous part, we use negative sampling to generate negative sharing user candidates uniformly. The prediction is based on similarity scores defined below

$$v_{v_d}^T \cdot (W_s v_s + b_s) \tag{9}$$

where v_d is the document vector for d, v_s the embedding of sharing user s, W_s and b_s the parameters for the hidden layer.

4.3 Frame Indicator Guided Pre-training

The frame indicator guided pre-training is almost identical to the entity guided one except that the masked tokens are frame indicators instead of entity mentions.

4.4 Ensemble of Multiple Models

Given the entity and user embeddings are not in the same space, we use them to pre-train separate models. All pre-trained models are then trained with the supervision of political perspective labels in the same way. We also explore an ensemble of the three models, which makes predictions based on a weighted sum of unnormalized scores f_d in equation 5 from these models at test time.

$$\sum_m f_{dm} * \beta_m \tag{10}$$

where m denotes a trained prediction model, f_{dm} the unnormalized scores for document d by model m and β_m the weight given to model m which can be tuned based on the data.

5 Experiments

We aim to answer the following research questions (RQs) in the experiment:

RQ1: what is the performance gain of pre-training the text model with each social and linguistic information, with respect to the baseline models?

RQ2: what is the respective contribution by the individual pre-trained models to the full ensemble model?

RQ3: how will the performance gain change given the different amount of labeled data available for training?

5.1 Datasets and Evaluation

We run experiments on two news article datasets: Allsides and SemEval. The statistics of both datasets are shown in Table 1.

Allsides This dataset (Li and Goldwasser, 2019) is collected from two news aggregation websites[4] on 2020 different events discussing 94 event types. The websites provide news coverage from multiple perspectives, indicating the bias of each article using crowdsourced and editorial reviewed approaches. Each article has a political perspective label left, center, or right. We used the same randomly separated splits for evaluation in this paper so that our results are directly comparable with previous ones.

SemEval This is the official training dataset from SemEval 2019 Task 4: Hyperpartisan News Detection (Kiesel et al., 2019). The task is to decide whether a given news article follows a hyperpartisan argumentation. There are 645 articles in this dataset and each is labeled manually with a binary label to indicate whether it is hyperpartisan or not. Since the test set is not available at this time. We conducted 10-fold cross-validation on the training set with the exact same splits so that we can compare with the system that ranked in the first place.

Dataset	Center	Left	Right	Avg # Sent.	Avg # Words
Allsides	4164	3931	2290	49.96	1040.05
		Hyperpartisan			
SemEval	407	238		27.11	494.29

Table 1: Datasets Statistics.

[4] Allsides.com and Memeorandum.com

5.2 Baselines

We compare our model with several competitive baseline methods.

BERT is a language representation model based on deep bidirectional Transformer architectures (Vaswani et al., 2017). It was pre-trained with the masked language model and next sentence prediction tasks on a huge corpus. As a result, it can achieve state-of-the-art results on a wide range of tasks by fine-tuning with just one additional output layer.

CNN_Glove (CNN_ELMo) is the model from the team that ranked first in hyperpartisan news detection task in SemEval 2019 (Jiang et al., 2019). It uses the pre-trained Glove (ELMo) word vectors, which are then averaged as sentence representations. The sentence vectors are fed into 5 convolutional layers of different kernel sizes. The outputs for all convolution layers are combined to form the input to a fully connected layer, which maps to the final text representation. Some extra improvements include batch normalization and ensemble of multiple models.

5.3 Implementation Details

We use the spaCy toolkit for preprocessing the documents. All models are implemented with PyTorch (Paszke et al., 2017)[5]. The 300d Glove word vectors (Pennington et al., 2014) trained on 6 billion tokens are used to convert words to word embeddings. The ELMo model we used is the medium one with output size 512. They are not updated during training. The sizes of LSTM hidden states for both word level and sentence level are 300 for both Allsides and SemEval datasets. The number of attention heads at both word and sentence levels is set to 4 for the Allsides dataset, while it is set to 1 for the SemEval dataset due to its size. For the training of the neural network, we used the Adam optimizer (Kingma and Ba, 2014) to update parameters. On Allsides dataset, 5% of the training data is used as the validation set. We perform early stopping using the validation set. However, same as (Jiang et al., 2019), we use the evaluation part of each fold for early stopping and model selection due to the limited size of the SemEval dataset. The patience for early stopping p is equal to 10, meaning that the training stops when there is no improvement in validation performance for ten consecutive epochs. The learning rate lr is set to 0.001 for all models except BERT for which $2e-5$ is used. The mini-batch size $b = 10$ for bias prediction.

Regarding pre-training data sources, we use the training set for Allsides, and extract 100,000 news articles for SemEval from the large dataset provided by SemEval 2019 Task 4. The entity and user embeddings used for pre-training are obtained through external resources described in Section 3.2 and 3.3. The embeddings for frame indicators are randomly initialized. All of them were updated during the pre-training to better adapt to the text model. The optimizer and most hyper-parameters stay the same as the training of bias prediction. The mini-batch size is set to 2000 and 300 for models using Glove and ELMo respectively since the training is at the sentence level.

5.4 Results

5.4.1 Results on Allsides

We report the average accuracy and macro F1 scores on test sets for Allsides dataset in Table 2. The results are divided into two groups based on whether contextualized word representations are used. To answer RQ1, we observed that, in most cases, models with pre-training outperform the MAN baseline. It demonstrates our pre-training step can effectively utilize signals in social and linguistic context to enhance the text model to identify bias expressed in more subtle ways. Therefore it generates high-quality document representation for political perspective prediction. The sharing guided pre-training did not lead to much improvement by itself. This is mainly because the sharing users in our dataset often share news articles with various perspectives. Our ensemble model achieves the best result in terms of both accuracy and macro F1 scores no matter whether contextualized word embeddings are used or not. It shows the signals from various sources are complementary with each other such that even a simple combination of prediction scores can lead to significant improvement. The gaps between our model and baselines decrease when contextualized word representations are used since local context is better captured in this setting.

5.4.2 Results on SemEval

The performance of various models on the SemEval dataset can be found in Table 3. Note that

[5]Please refer to https://github.com/BillMcGrady/NewsBiasPretraining for data and source code.

Model	Accuracy	Macro F1
MAN_Glove	78.29	76.96
+ Entity	80.50	79.50
+ Sharing	78.93	77.84
+ Frame	81.26	80.15
Ensemble	**83.74**	**82.84**
BERT	81.55	80.13
MAN_ELMO	81.41	80.44
+ Entity	82.27	81.23
+ Sharing	81.37	80.48
+ Frame	82.56	81.66
Ensemble	**85.00**	**84.25**

Table 2: Test Results on Allsides Dataset.

there is no sharing user guided result in this table since we do not have social graph information available in this dataset. Again the results are grouped based on word representation used. CNN_Glove and CNN_ELMo are results reported by the winning team in the SemEval competition. They proposed an ensemble of multiple CNN models where each CNN takes sentence representation generated by average ELMo embedding as input. It is worth noting that our model with Glove as word representation is comparable with the winning team's model with ELMo, showing the advantages of pre-training. The other trends hold as well in the SemEval dataset. In both datasets, our pre-trained models beat BERT easily since they are tuned specifically for the task.

Model	Accuracy	Macro F1
CNN_Glove †	79.63	-
MAN_Glove	81.58	79.29
+ Entity	82.65	80.75
+ Frame	83.27	81.73
Ensemble	**84.03**	**82.42**
CNN_ELMO †	84.04	-
BERT	84.03	82.60
MAN_ELMO	84.66	83.09
+ Entity	85.59	84.15
+ Frame	85.27	83.32
Ensemble	**86.21**	**84.33**

Table 3: Test Results on SemEval Dataset. † indicates results reported in (Jiang et al., 2019).

5.4.3 Ablation Study

To answer RQ2, we show the results for ablations of our ensemble model based on MAN_Glove in Table 4. The performance drops when removing each one of the pre-trained models from the ensemble, showing that the information obtained from different sources is complementary with each other. To make a fair comparison with the baseline model, we also report the performance of an ensemble of multiple baseline models (denoted as -Pre-training) with different seeds from random initialization. This shows the absolute gain through pre-training to adapt the text representations for political perspective identification.

Model	Accuracy	Macro F1
Ensemble	83.74	82.84
- Entity	82.57	81.65
- Sharing	82.78	81.78
- Frame	82.39	81.40
- Pre-training	81.54	80.40

Table 4: Ablation Study on Allsides Dataset.

5.4.4 Results with Limited Training Data

One of the obstacles in obtaining good performance in political perspective identification tasks is the lack of supervision data. We compare the performance of the MAN_Glove model with and without pre-training with different levels of training examples available in Figure 2. These results can help to answer RQ3. It shows that the performance gain obtained from our pre-training strategy increases as the size of the training set decreases. This is a very useful property as it can greatly improve model performance when there is limited training data. It is worth noting that the Sharing-Guided Pre-training achieves much higher performance when supervision is limited. This is because the signals from the sharing users can be considered as noisy bias labels and it is trained at document level instead of sentence level like the other two. However, since the other two pre-training methods introduce extra knowledge to the text model, they can lead to better performance when the supervision is abundant to provide enough bias information for training.

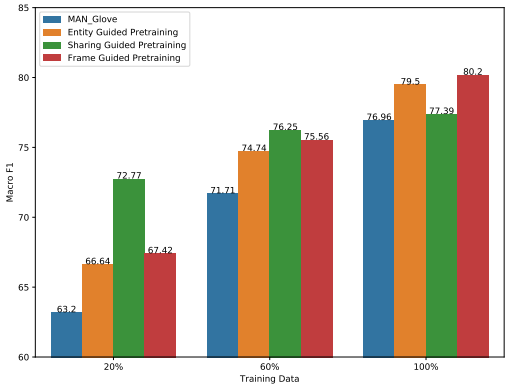

Figure 2: Test Results with Different Number of Training Examples.

5.4.5 Qualitative Results

Human Annotation Comparison The BASIL dataset (Fan et al., 2019) has human annotations of bias spans. It contains 300 articles on 100 events with 1727 bias spans annotated. On the sentence level, spans of lexical and informational bias are identified by annotators by analyzing whether the text tends to affect a reader's feeling towards one of the main entities. We compute the average attention assigned by our model to the annotated bias spans. Table 5 shows the results of the baseline model (MAN) and the same model pre-trained with entity information (+Entity). The attention scores assigned to the human annotation spans are higher across training, validation, and test sets.

Model	Training	Validation	Test
MAN	0.706	0.701	0.652
+ Entity	0.737	0.728	0.666
Improvement	4.36%	3.76%	2.13%

Table 5: Average Attention Scores on Basil Annotations.

6 Conclusion

In this work, we propose a pre-training framework to adapt text representation for political perspective identification. Empirical experiments on two recent news article datasets show that an ensemble of pre-trained models achieves significantly better performance in bias detection compared to competitive text baselines. It is also shown that our pre-training model can achieve even larger performance gain when the supervision is limited.

In fact, these various context information are not independent. We intend to extend this work to pre-train better text models by incorporating information from various sources together.

Acknowledgements

We thank the anonymous reviewers for their helpful feedback. This works was supported by an NSF CAREER award IIS-2048001.

References

E. Baumer, E. Elovic, Y. Qin, F. Polletta, and G. Gay. 2015. Testing and comparing computational approaches for identifying the language of framing in political news. In *NAACL*, pages 1472–1482, Denver, Colorado. Association for Computational Linguistics.

C. Budak, S. Goel, and J. M. Rao. 2016. Fair and balanced? quantifying media bias through crowdsourced content analysis. *Public Opinion Quarterly*, 80(S1):250–271.

Dallas Card, Justin Gross, Amber Boydstun, and Noah A. Smith. 2016. Analyzing framing through the casts of characters in the news. In *EMNLP*, pages 1410–1420, Austin, Texas. Association for Computational Linguistics.

J. Daiber, M. Jakob, C. Hokamp, and P. N. Mendes. 2013. Improving efficiency and accuracy in multilingual entity extraction. In *I-SEMANTICS*, I-SEMANTICS '13, pages 121–124, New York, NY, USA. ACM.

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

Heba Elfardy, Mona Diab, and Chris Callison-Burch. 2015. Ideological perspective detection using semantic features. In *STARSEM*, pages 137–146, Denver, Colorado. Association for Computational Linguistics.

L. Fan, M. White, E. Sharma, R. Su, P. K. Choubey, R. Huang, and L. Wang. 2019. Media bias through the lens of factual reporting. In *EMNLP*.

A. Field, D. Kliger, S. Wintner, J. Pan, D. Jurafsky, and Y. Tsvetkov. 2018. Framing and agenda-setting in Russian news: a computational analysis of intricate political strategies. In *EMNLP*, pages 3570–3580, Brussels, Belgium. Association for Computational Linguistics.

M. Gardner, J. Grus, M. Neuman, O. Tafjord, P. Dasigi, N. F. Liu, M. Peters, M. Schmitz, and L.S. Zettlemoyer. 2017. Allennlp: A deep semantic natural language processing platform.

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

Matthew Gentzkow and Jesse M Shapiro. 2011. Ideological segregation online and offline. *The Quarterly Journal of Economics*, 126(4):1799–1839.

Stephan Greene and Philip Resnik. 2009. More than words: Syntactic packaging and implicit sentiment. In *NAACL-HLT*, pages 503–511, Boulder, Colorado. Association for Computational Linguistics.

K. Hanawa, S. Sasaki, H. Ouchi, J. Suzuki, and K. Inui. 2019. The sally smedley hyperpartisan news detector at SemEval-2019 task 4. In *SemEval*, pages 1057–1061, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. 2014. Political ideology detection using recursive neural networks. In *ACL*, pages 1113–1122, Baltimore, Maryland. Association for Computational Linguistics.

Ye Jiang, Johann Petrak, Xingyi Song, Kalina Bontcheva, and Diana Maynard. 2019. Hyperpartisan news detection using ELMo sentence representation convolutional network. In *SemEval*, pages 840–844, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Kristen Johnson and Dan Goldwasser. 2016. Identifying stance by analyzing political discourse on twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 66–75.

Kristen Johnson, Di Jin, and Dan Goldwasser. 2017. Modeling of political discourse framing on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11.

J. Kiesel, M. Mestre, R. Shukla, E. Vincent, P. Adineh, D. Corney, B. Stein, and M. Potthast. 2019. SemEval-2019 task 4:hyperpartisan news detection. pages 829–839, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

C. Li and D. Goldwasser. 2019. Encoding social information with gcn for political perspective detection in news media. In *ACL*, pages 2594–2604, Florence, Italy. Association for Computational Linguistics.

W-H Lin, T. Wilson, J. Wiebe, and A. Hauptmann. 2006. Identifying perspectives at the document and sentence levels. In *CoNLL*, CoNLL-X '06, pages 109–116, Stroudsburg, PA, USA. Association for Computational Linguistics.

Julia Mendelsohn, Ceren Budak, and David Jurgens. 2021. Modeling framing in immigration discourse on social media. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2219–2263.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

F. Morstatter, L. Wu, U. Yavanoglu, S. R. Corman, and H. Liu. 2018. Identifying framing bias in online news. *Trans. Soc. Comput.*, 1(2):5:1–5:18.

Van-Hoang Nguyen, Kazunari Sugiyama, Preslav Nakov, and Min-Yen Kan. 2020. Fang: Leveraging social context for fake news detection using graph representation. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 1165–1174.

Maria Leonor Pacheco and Dan Goldwasser. 2021. Modeling content and context with deep relational learning. *Transactions of the Association for Computational Linguistics*, 9:100–119.

A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. 2017. Automatic differentiation in PyTorch. In *NIPS Autodiff Workshop*.

J. Pennington, R. Socher, and C. Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

M. Recasens, C. Danescu-Niculescu-Mizil, and D. Jurafsky. 2013. Linguistic models for analyzing and detecting biased language. In *ACL*, pages 1650–1659, Sofia, Bulgaria. Association for Computational Linguistics.

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

Chereen Shurafa, Kareem Darwish, and Wajdi Zaghouani. 2020. Political framing: Us covid19 blame game. In *International Conference on Social Informatics*, pages 333–351. Springer.

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 I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Hongwei Wang, Fuzheng Zhang, Xing Xie, and Minyi Guo. 2018. Dkn: Deep knowledge-aware network for news recommendation. In *WWW*, WWW' 18, pages 1835–1844, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

J. Wang, Z. Wang, D. Zhang, and J. Yan. 2017. Combining knowledge with deep convolutional neural networks for short text classification. In *IJCAI*, IJCAI, pages 2915–2921. AAAI Press.

Hu Xu, Bing Liu, Lei Shu, and Philip Yu. 2019. BERT post-training for review reading comprehension and aspect-based sentiment analysis. 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 2324–2335, Minneapolis, Minnesota. Association for Computational Linguistics.

I. Yamada, A. Asai, H. Shindo, H. Takeda, and Y. Takefuji. 2018. Wikipedia2vec: An optimized tool for learning embeddings of words and entities from wikipedia. *arXiv*.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California. Association for Computational Linguistics.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy. Association for Computational Linguistics.