POLOR: Leveraging Contrastive Learning to Detect Political Orientation of Opinion in News Media

Ala Jararweh
The University of New Mexico
Albuquerque NM, USA
ajararweh@unm.edu

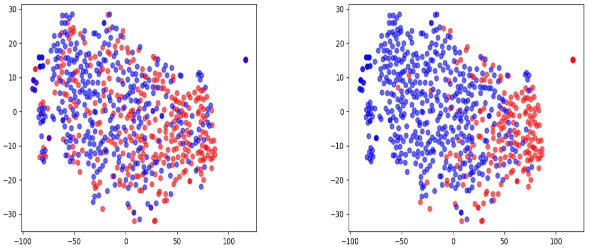
Abdullah Mueen
The University of New Mexico
Albuquerque NM, USA
mueen@unm.edu

Abstract

News articles are naturally influenced by the values, beliefs, and biases of the reporters preparing the stories and the policies of the publishing outlets. Numerous studies and datasets have been proposed to detect the political orientation of news articles. However, most of these studies ignore real textual clues and learn the textual signature of the source (commonly the publisher and rarely the writer) of the article instead. Moreover, a good volume of opinion pieces published by major news outlets do not reflect the political orientation of the publisher but rather reflect the political orientation of a non-professional writer. Existing methods are not built to correct this difference in the training data and, hence, perform poorly on human-annotated data. We propose, POLOR, a fine-tuned BERT model that employs contrastive learning to detect the political orientation of news articles even when the training data is labeled by the source (i.e. the publisher of the news article). Unlike previous work in the literature, the model learns features by employing different contrastive learning objectives where each sentence is contrasted with sentences from various sources simultaneously. POLOR achieves a 15% increase on our dataset compared to previously proposed baselines. Finally, we release two datasets of opinion news: source-annotated and human-annotated datasets. The code and datasets can be found at `https://www.cs.unm.edu/~ajararweh/`.

1 Introduction

Political bias in news can be emphasized in various forms such as rephrasing, or presenting one-sided facts to serve a specific orientation. This bias has a direct impact on the information consumed by the readers and the political attitudes (Gentzkow and Shapiro 2006). Detecting the political orientation of news is a challenging task due to the necessity of understanding the multidimensional political discourse. For example, mainstream media worldwide use terms like ”explosion” versus ”attack” when reporting on recent conflicts in the Middle East to favor one political position over another (de Jong 2023). This issue has been studied in various contexts and a variety of datasets has been proposed. Reserchers use online platforms such as `AllSides`[1], `AdFontesMedia`[2], `MediaBiasFactCheck`[3], `NewsGuard`[4], etc to collect and annotate news articles.



(a) Crowd-Sourcing Labels

(b) Source Labels

Figure 1: Visualized article embeddings generated by POLOR, and annotated according to (a) article orientation, and (b) source orientation. The blue points represent *Liberal*, and the red points represent *Conservative*. Source annotations do not always reflect all article orientations published by that source. Consequently, models trained on source annotations without utilizing auxiliary information, learn to predict sources instead of capturing article-specific orientations in Figure 1(a).

Such platforms usually offer political annotations at the news source level. To obtain article-specific orientation, annotations are usually propagated from the source annotations (Lee et al. 2022; Baly et al. 2020; Chen et al. 2018; Liu et al. 2022; Kulkarni et al. 2018). For example, `AllSides.com` annotates all articles published by the New York Times as Liberal. Figure 1 shows visualized embeddings of articles generated by our model. The dimensions of the embeddings were reduced using t-distributed Stochastic Neighbor Embedding (van der Maaten and Hinton 2008). The true article orientations (Figure 1(a)) differ from the annotations de-

[1] `https://www.allsides.com`
[2] `https://adfontesmedia.com`
[3] `https://mediabiasfactcheck.com`
[4] `https://www.newsguardtech.com`

rived from sources (Figure 1(b)). The red points are more distributed to the right and the blue points are to the left in both figures which illustrates that the source orientations can be related to the article orientations.

Relying on the source annotation without utilizing auxiliary knowledge, the trained models may attempt to learn the writing style of the news source instead of predicting the orientations of news articles. The problem is emphasized when opinion pieces are considered. Opinion pieces usually represent the bias of the authors, and are often published by a media with opposite bias. This problem has been lightly touched on by Baly et al. (2020), where the authors found that the model suffers when trained on source-labeled data and tested on unseen data (i.e. new media sources). To overcome this, they inserted a source classifier to minimize the knowledge learned from the source. However, they have not tested their model on human-annotated data where the labels are derived based on the text in articles, paragraphs, or sentences.

In this paper, we propose **POLOR**: leveraging contrastive learning to detect the **POL**itical **OR**ientation of opinion pieces in news media. POLOR exploits several contrastive learning objectives where each sentence is contrasted to a set of sentences from different and similar sources. We adopt different objective functions to generate additional features that focus on textual cues related to the political bias of articles, instead of the style of the news media. POLOR produces sentence-specific and article-specific labels based on the training data derived from sources.

2 Related Work

Detecting the political orientation in news media can be challenging due to the necessity of domain expertise. In their early attempts, researchers measured the orientation of a particular news media by counting the number of citations to liberal and conservative think-tank policy groups (Groseclose and Milyo 2005). Focusing on the article content, Greene and Resnik (2009) used domain-specific words in news articles to classify articles that support Palestine or Israel. Similarly, researchers harnessed the recent findings in Deep Learning to understand the political orientation of news articles at different levels of article structure such as character-based word representations (Jiang et al. 2019), sentences (Kim and Johnson 2022; Gangula, Duggenpudi, and Mamidi 2019; Chen et al. 2018), and articles (Baly et al. 2020; Chen et al. 2020; Liu et al. 2022).

Various datasets were proposed to tackle this task with a variety of annotation techniques. Since it's challenging and expensive to annotate a large corpus of news articles, different studies rely on online platforms such as `AllSides.com` to crawl and annotate news articles (Baly et al. 2020; Chen et al. 2020; Liu et al. 2022; Roy and Goldwasser 2020). The annotations of articles are usually propagated from the source orientation. However, a major drawback of these annotations is that models learn to predict the source signature rather than the article's orientation. To overcome this, researchers use online crowd-sourcing platforms to obtain article-specific annotations (Gangula, Duggenpudi, and Mamidi 2019; Fan et al. 2019; Card et al. 2015; Budak, Goel, and Rao 2016; Lazaridou et al. 2020). To address the problem of predicting article-specific labels from source-based annotations, few datasets were found that integrate both annotations. Kiesel et al. (2019) proposed two datasets but the human annotation dimensions of articles are different from our task. To this end, we construct a new dataset that spans 10 news sources with two different annotations.

	Liberal	Conservative	
Articles#	156	107	263
paragraphs#	5146	3099	8245
Sentences#	7743	3920	11921
Sources#	4	3	7

(a) Statistics of the Source-Based dataset.

	Liberal	Conservative	Neutral	
Articles#	22	17	0	39
Paragraphs#	138	133	42	313
Sentences#	522	513	188	1223
Sources#	-	-	-	3

(b) Statistics of the Human-annotated dataset.

Table 1: Exploring the constructed dataset for this task.

3 Datasets

We evaluate our model performance on two datasets. The first dataset consists of approximately 6,000 articles annotated via crowd-sourcing from Budak, Goel, and Rao (2016). It also spans 11 news sources and explores 15 different topics. Since the source annotations are not available, we extend the dataset by deriving annotations from `AllSides.com`. For completeness, we further discuss the dataset in Appendix C. For a broader evaluation, we construct another dataset for this task. Our dataset consists of two subsets, crowd-sourcing annotations and source annotations. The constructed dataset addresses two high-profile criminal cases, namely *Ahmaud Arbery* and *Kyle Rittenhouse*. A full description of the cases can be found in Appendix A.

Source-Based Dataset. We collected a total of 263 news articles which cover 7 different U.S. news sources. We align with recent work in the literature, we obtained annotations from `AllSides.com`. This website only offers annotations at the source level based on multiple methods such as Editorial Review, Blind Bias Survey, and Community Feedback. We collect articles from 4 liberal news sources and 3 conservative news sources according to the annotations found on the website. Table 1(a) shows some detailed statistics of the dataset. We follow the same annotation process of recently published datasets where the article annotations are propagated from their source annotations.

Human-Annotated Dataset. We also construct a smaller dataset for evaluating the model on human annotations. This dataset consists of 39 news articles from three alternative U.S. news sources, different from those in the training data

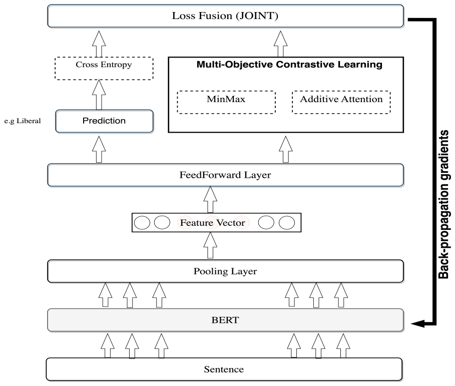


Figure 2: An overview of POLOR, which utilizes contrastive learning objectives to extrapolate article-specific labels.

(Source-Based Dataset). We aim to evaluate our model's ability to generalize to new writing styles. Thus, we only consider opinion news published about the two cases. The articles were collected via word matching from the news media websites. The annotations were performed at the paragraph level, where each article was partitioned into a set of paragraphs to provide annotators with more context about the cases. The article annotations are then derived based on majority voting. Each paragraph is labeled by three different workers into one of three different categories: Liberal, Conservative, or Neutral. The workers were provided with a set of instructions and facts about the two cases ahead of time. 92% of the articles received an inter-annotator score above 0.9, and 8.0% are below 0.9. The articles that have low scores (below 0.5) were annotated Neutral. The news source of each paragraph was hidden from the workers and they only could see the content of the paragraph. Table 1(b) shows the detailed statistics and the annotation results. We provide a detailed description of this dataset in Appendix B.2. To match with annotation dimensions in the source data, we only consider conservative and liberal paragraphs in our evaluation.

Data Preparation. The datasets are further pre-processed by removing unnecessary content such as punctuation, non-English characters, and identifying words and sentences. We then split each article/paragraph into a set of sentences. The final preparation step is to formulate the datasets as triplets. Each sentence in the dataset S_a (also called Anchor) is attached to a set of positive sentences S_p, and a set of negative sentences S_n. We define a sentence as positive if it belongs to the same class and from a different source, and negative if it belongs to the opposite class. For example, an anchor labeled as ”Conservative”, its positive set is randomly derived from other sentences labeled as ”Conservative”, and its negative set is randomly derived from sentences labeled as ”Liberal”. We also ensured that the triplets of the training, validation, and test set were completely disjoint. The size of the positive and negative sets ($\mathcal{T}$) is a parameter.

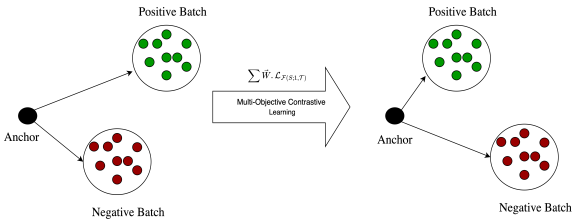

Positive Batch
Anchor
Negative Batch
$\sum \vec{W} . \mathcal{L}_{\mathcal{F}(S;1,\mathcal{T})}$
Multi-Objective Contrastive Learning
Positive Batch
Anchor
Negative Batch

Figure 3: The desired goal of Multi-objective Contrastive Learning.

4 Methodology

We propose POLOR, a complete framework that leverages multi-objective contrastive learning to generate embeddings that portray the bias in news articles by minimizing the source information. Figure 2 shows the main components of the model, namely Sentence Embedding, Multi-Objective Contrastive Learning, Prediction, and Loss Fusion (Joint Loss).

4.1 Sentence Embedding

The sentences are first tokenized, where each sentence in anchor, negative, or positive sets is represented as a list of N tokens. The tokens are passed to the pre-trained BERT model to generate contextualized token representations (i.e. embeddings) where each token, T_i, is represented in an embedding vector, E_i. Figure 2 shows the input and output of BERT. At the end of this stage, each sentence S will be represented with N contextualized token representations, where $E_i \in \mathbb{R}^{768}$. To synthesize sentence embeddings, we use the pooling strategies, CLS (classification token) Pooling and Mean Pooling introduced by Reimers and Gurevych (2019).

4.2 Multi-Objective Contrastive Learning

Our goal is to tune the representations such that they exclusively represent the bias in the content of articles. Since the training dataset (Source-Based Dataset) only contains sentences annotated based on sources, the model learns additional features in an unsupervised manner. That is, the source influence is minimized and the influence of the political orientation in the text is maximized. To achieve this goal, we employ different objective functions that utilize contrastive learning. In the results section, we empirically show that augmenting contrastive objectives on top of BERT helps to learn rich features by contrasting other sentences from similar/different sources in the dataset. Figure 3 demonstrates the desired goal of using multi-objective contrastive learning where we attempt to find a better representation by employing a triplet loss function. The triplet loss function was introduced by Schroff, Kalenichenko, and Philbin (2015) in the field of face recognition to optimize face image embeddings. The authors define positive examples as similar images of the same face, and negative examples are all other faces in the mini-batch.

We adopt an alternative solution to this. We first assign a set of $\mathcal{T}$ positives and a set of $\mathcal{T}$ negatives from the entire dataset at random for each anchor. We then apply different objective functions on the positive and negative sets to synthesize rich representations to achieve different optimizations. After applying each objective function, the synthesized representation should have a similar dimension to the anchor ($\mathbb{R}^{768}$). The synthesized representations of positive and negative sets are then contrasted with the anchor to obtain the loss for the desired objective. Given an anchor sentence S_a, set of positive sentences S_p and, and set of negative sentences S_n, where S_p and S_n of size $\mathcal{T}$, we re-frame the loss function as follows:

$$\mathcal{L}(S_a, S_p, S_n) = \\ max \begin{cases} \|S_a - \mathcal{F}(S_p; 1, \mathcal{T})\|_2 - \|S_a - \mathcal{F}(S_n; 1, \mathcal{T})\|_2 + m \\ 0 \end{cases} \tag{1}$$

where m is the margin to be enforced between the positive and negative sets, and $\mathcal{F}(S; 1, \mathcal{T})$ is an objective function used to synthesize a representation from the set of positive and negative sets. In the following subsections, we describe the objective functions used to synthesize the representations. We then define the Joint loss function which is designed to fuse all losses, obtained by contrasting data to the synthesized representations, into one universal function.

Additive Attention. Since the sentences are annotated based on source, the annotation does not always reflect the true orientation of the sentences. The model attends to the sentences in the positive and negative sets based on importance and relativity to the anchor. We modify the additive attention mechanism mentioned in Bahdanau, Cho, and Bengio (2014) to serve our goal. This attention mechanism was mainly proposed in machine translation to synthesize a context vector of words in a sentence. The context vector is a weighted sum of the encoder's hidden state with the attention scores. The attention scores are usually computed by feeding the concatenation of the encoder's current hidden state and the decoder's previous hidden state to an alignment model (a non-linear transformation of the input).

In our case, the position of sentences in the positive/negative set does not matter in predicting the anchor. Consequently, we replace the hidden state of the encoder and the decoder with two feed-forward layers where the input is the positive/negative sentence embeddings. The results are then concatenated and fed to an alignment model to calculate the attention scores.

$$\begin{aligned} e &= W_3.tanh([W_1 \cdot E; W_2 \cdot E]) \\ a_i &= \frac{exp(e_i)}{\sum_{k=1}^{\mathcal{T}} exp(e_k)} \end{aligned} \tag{2}$$

where E is the embedding of sentences from the positive/negative set. The final step is to calculate the weighted sum of positive/negative embeddings with the attention scores:

$$\mathcal{F}(E; 1, \mathcal{T}) = \sum_{i=1}^{\mathcal{T}} a_i.E_i \tag{3}$$

where we obtain $\mathcal{L}_{\mathcal{F}_{additive}}$ by substituting $\mathcal{F}(E; 1, \mathcal{T})$ in Eq. 1 for the negative and positive sets separately.

Unsupervised MinMax. The goal of this objective function is to find the appropriate negative and positive sentences extracted from the entire positive and negative sets. We define a sentence as appropriate such that it influences the model to increase the margin between the negative set and the positive set in the Euclidean space. To achieve this goal, unsupervised MinMax treats the positive and negative sentences as one set where the closest sentence is considered negative and the furthest sentence is considered positive.

$$\begin{aligned} \mathcal{F}(S_p; 1, \mathcal{T}) &= \text{argmax}_{s \in (S_p \cup S_n)} \|g(S_a) - g(s)\|_2^2 \\ \mathcal{F}(S_n; 1, \mathcal{T}) &= \text{argmin}_{s \in (S_p \cup S_n)} \|g(S_a) - g(s)\|_2^2 \end{aligned} \tag{4}$$

where $g(s)$ is the embedding of the sentence s. The loss function $\mathcal{L}_{\mathcal{F}_{MinMax}}$ is then obtained by substituting in Eq. 1. This selection ensures fast convergence and helps to increase the margin between the anchor and negative sentences and decrease it with positive sentences (Schroff, Kalenichenko, and Philbin 2015).

4.3 Joint Loss (Fusion)

The losses obtained by the above objective functions are assembled to allow the model to learn better representation. We use different hyper-parameters to weigh the objective losses and then join them into one loss function. In addition to the unsupervised contrastive loss, we incorporated Binary Cross-Entropy loss ($\mathcal{L}_{XEnt}$) in the final loss function to obtain the Join loss ($\mathcal{L}_J$):

$$\begin{aligned} \mathcal{L}_J = \alpha \cdot \mathcal{L}_{\mathcal{F}_{additive}} + \beta \cdot \mathcal{L}_{\mathcal{F}_{MinMax}} \\ + (1 - \alpha - \beta) \cdot \mathcal{L}_{XEnt} \end{aligned} \tag{5}$$

where $\mathcal{L}_{\mathcal{F}}$ is the triplet loss obtained by applying a specific objective function, and $0 < \alpha, \beta < 0.5$. Note that, when a parameter equals 0, the associated loss function is not incorporated into the Joint loss.

5 Experiments and Results

5.1 Baselines

Our goal is to classify the article-based political orientation of an article using training data labeled by their sources. We evaluate our model performance against different baselines with different training setups of our model:

1. **Majority classifier**: We consider the most frequent class in the dataset as the predicted label for all sentences.
2. **Joint function with XEnt loss only** : We use the sentence representations generated by BERT to perform classification in a supervised manner. For this baseline, the contrastive learning loss is entirely ignored (i.e. $\alpha = 0$ and $\beta = 0$). This approach is similar to Fan et al. (2019).
3. **Joint function with XEnt and Additive**: We considered two loss functions in the joint function (Eq. 5), namely additive and XEnt. That is, β is set 0.

		Large Media		Our dataset			
Losses	Model	Acc$_{Art}$	F1$_{Art}$	Acc$_{Sent}$	F1$_{Sent}$	Acc$_{Art}$	F1$_{Art}$
Majority	-	0.52	0.34	0.50	0.34	0.51	0.34
XEnt (BERT)	Fan et al. (2019)	0.63	0.60	0.51	0.51	0.62	0.61
XEnt + Additive	ours	0.65	0.62	0.53	0.53	0.67	0.65
XEnt + MinMax	ours	0.63	0.60	0.53	**0.53**	0.69	0.68
XEnt + Random	Kim and Johnson (2022), Baly et al. (2020)	0.62	0.59	0.53	0.47	0.64	0.64
XEnt + Centroid	ours	0.63	0.61	0.53	0.53	0.64	0.62
XEnt + Additive + MinMax	POLOR	**0.68**	**0.66**	**0.55**	0.520	**0.74**	**0.74**

Table 2: The results of comparing our multi-objective model (POLOR) with the baselines. The results on our dataset are reported on sentence predictions (Sent) and article predictions (Art) using accuracy (Acc) and macro F1 score metrics. For the large media study, the model was directly trained on articles.

4. **Joint function with XEnt and MinMax**: We consider two loss functions in the joint function (Eq. 5), namely XEnt and MinMax. That is, β is set 0.

5. **Joint function with XEnt loss and Random Triplet objective**: We consider only one contrastive objective function that selects a random positive and negative sentence from the sets S_p and S_n at each iteration. The joint loss becomes:

$$\mathcal{L}_J = \alpha \cdot \mathcal{L}_{\mathcal{F}_{RT}} + (1 - \alpha) \cdot \mathcal{L}_{XEnt}$$

   This setup is approximately similar to Kim and Johnson (2022) and Baly et al. (2020) with few differences. In their formulation, the triplet is usually chosen at random in an offline manner (i.e. when the dataset is created) and remains fixed throughout the whole training process. Moreover, their method of choosing positive and negative sentences is different from our approach. For example, Kim and Johnson use predefined sub-framing groups to assign the triplets, while Baly et al. use media sources.

6. **Joint function with XEnt loss and Centroid Triplet**: We consider only one contrastive objective function which is the mini-batch mean. In this objective, the positive is the centroid of the positive set and the negative is the centroid of the negative set. The joint loss becomes:

$$\mathcal{L}_J = \alpha \cdot \mathcal{L}_{\mathcal{F}_{centroid}} + (1 - \alpha) \cdot \mathcal{L}_{XEnt}$$

Parameter	**POLOR dataset**	**Large Media**
Model	bert-base-uncased	all-distilroberta-v1
α	0.14	0.29
β	0.16	0.13
Margin(m)	1.9	0.57
LR	$6.8e^{-4}$	$8.3e^{-5}$
Pooling	Mean	CLS
$\mathcal{T}$	5	5

Table 3: The best-performing hyper-parameters found for the two datasets

5.2 Model Configurations

We train our model on sentences where we use the pretrained BERT variants (bert-base-uncased and bert-base-cased) or sentence-transformers based models (all-mpnet-base-v2, all-MiniLM-L12-v1, all-MiniLM-L12-v2, and all-distilroberta-v1) as the encoder. All of the aforementioned encoders are available on Huggingface. The input is tokenized, truncated, or padded to a specific length (`maxlen`). The `maxlen` is varied between (80-250) [5]. We used `AdamW` optimizer with a learning rate between ($1e^{-4}$ - $8e^{-6}$), batch size between (80-125) [5], number of epochs between (3-8), and gradient clippingequals 1. For the joint loss parameters, we varied the value of α and β such that their upper limits sum to 1 (i.e. 0-0.5). For the mini-batch triplet size ($\mathcal{T}$), we varied this between (1-5) [5]. The `margin` (m) of the triplet loss function is also varied between (0-5).We trained the models on 4 Nvidia Tesla V100 SXM2 with 32GB memory GPU each. The time for an epoch depends on the triplet size used during the training, but on average it takes around 5 minutes. The best-performing model parameters for both datasets can be found in Table 3.

5.3 Quantitative Results

Our Dataset. Table 2 shows the performance of our model in predicting political orientation compared to the proposed baselines. POLOR outperforms all baseline models on sentence accuracy, article accuracy, and article macro F1 score. We use majority voting (highest occurring class) on sentences to predict the article labels. We first notice that the baseline models outperform the majority classifier when predicting the orientation of articles, but struggled with sentence orientation. Because sentence labels are propagated from articles, which are subsequently propagated from the sources. Ideally, sentence-specific labels would solve the problem. However, in the absence of sentence-specific labels, article-specific labels from independent sources would improve the performance. Moreover, choosing the proper objective function can drastically affect performance. Sophisticated objective functions may not always yield the de-

[5] We varied this parameter according to the machine capabilities.

sired performance. For example, the simple random triplet objective shows a competitive performance even though it requires almost no additional computations beyond selecting a sentence at random. Finally, additive and MinMax achieve a good performance, and when combined (POLOR), they achieve the highest outcomes. The combined objectives also illustrate a good performance on unseen source signatures, since the news sources in the testing set were not seen during training.

Large Media Dataset. The main objective of this experiment is to analyze the model behavior on a larger dataset. We evaluate our model performance on a dataset of 6089 news articles from Budak, Goel, and Rao (2016). The dataset is only annotated via crowd-sourcing. To obtain source annotations, we manually annotate articles based on source annotations from `ALLSIDE.com` website. We test our model performance on the entire article text to reduce the runtime. Table 2 shows the experiment results on this dataset. Our model of the combined Additive and MinMax objectives is still outperforming the baselines by a good margin. We also note that the additive objective separately achieves a competitive performance to POLOR.

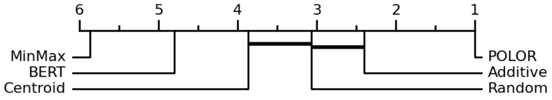


Figure 4: Critical difference diagram showing pairwise statistical difference comparison of the proposed baselines.

Significance Test. To further explain POLOR performance compared to the proposed baselines, we conduct a critical difference test based on the Wilcoxon-Holm method to detect pairwise significance. The test was performed using 15-fold cross-validation on the large media dataset only. We have not performed the test on our dataset since the annotated test set is fixed, and running the experiment k-times would always yield the same results. Figure 4 shows the results of the significance test with a significance level of 0.05. POLOR ranks first compared to the proposed baselines which demonstrated a significant performance in predicting human-level labels from news source annotations. The test also illustrates the superior capability of the additive objective in this task, unlike the MinMax objective which is anticipated to be impacted by the distances between the articles in the positive and negative sets.

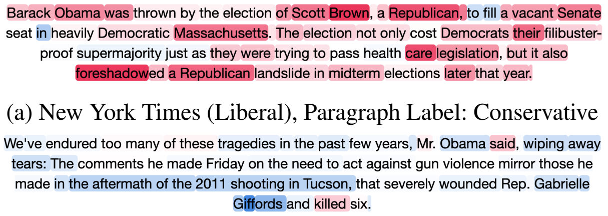
Parameter Sensitivity Study The choice of values alpha and beta can drastically affect the model performance. Table 4 shows the model performance on article accuracy for various alpha and beta values. Higher values of alpha and beta lead to poor performance since the cross-entropy loss (XEnt) value will be near zero which means that the prediction will be entirely unsupervised. On the other hand, smaller values of alpha and beta can yield good performance. Moreover, the model is also sensitive to the choice of MinMax margin. Enforcing large margins between anchors and their negative and positive sets does not always attain beneficial outcomes. Table 5 shows the performance of POLOR on different margins while other hyper-parameters are fixed. Large and small values can negatively affect the efficiency of the model while margins centered around 2, yield the desirable performance.

β \ α	0.136	0.424	0.5
0.157	0.743	0.6666	0.564
0.177	0.564	0.692	0.589
0.328	0.538	0.513	0.538

Table 4: The effect of α and β values on the article accuracy.

Margin	Acc_{Ar}	Acc_{Snt}
0.000	0.513	0.501
0.789	0.564	0.522
1.919	0.743	0.553
5.000	0.513	0.502

Table 5: The effect of MinMax margin on POLOR performance.


Barack Obama was thrown by the election of Scott Brown, a Republican, to fill a vacant Senate seat in heavily Democratic Massachusetts. The election not only cost Democrats their filibuster-proof supermajority just as they were trying to pass health care legislation, but it also foreshadowed a Republican landslide in midterm elections later that year.

(a) New York Times (Liberal), Paragraph Label: Conservative

We've endured too many of these tragedies in the past few years, Mr. Obama said, wiping away tears: The comments he made Friday on the need to act against gun violence mirror those he made in the aftermath of the 2011 shooting in Tucson, that severely wounded Rep. Gabrielle Giffords and killed six.

(b) Wall Street Journal (Conservative), Paragraph Label: Liberal

Figure 5: Visualizing the model attention on two real examples where the text label contradicts the source label.

5.4 Qualitative Results

Visual Analysis of Model Attention. We further demonstrate the model performance by delving deeper into two correctly predicted sample paragraphs where the source annotations contradict the text annotations. Figure 5 depicts the model attention using the SHAP explainer (Lundberg and Lee 2017) where the conservative and liberal contributions are reflected by the red and blue highlights respectively. Although the selected examples are intricate and potentially elusive for some readers to catch the subtle undertone, the model efficiently allocated more attention to underlying signaling words and phrases. In Figure 5(a), the model demonstrates higher attention toward the term `"Republican"` compared to the term `"Democrats"`. Similarly, the phrase `"their ... care legislation"` also contributes to the paragraph prediction since it is proposed or supported by a democratic group as the paragraph states, and the phrase `"foreshadowed a Republican"` which indicates a positive framing of anticipated Republican success. For Figure 5(b), the model attended to framing words that tend to humanize former President Barack Obama such as "wiping away tears" where it falls short in capturing other phrases like "act against gun violence". Since the paragraph labels exclusively contradict the source labels, the given remarks qualitatively suggest that the model devotes more attention to the signaling phrases in the text rather than learning the source writing style.

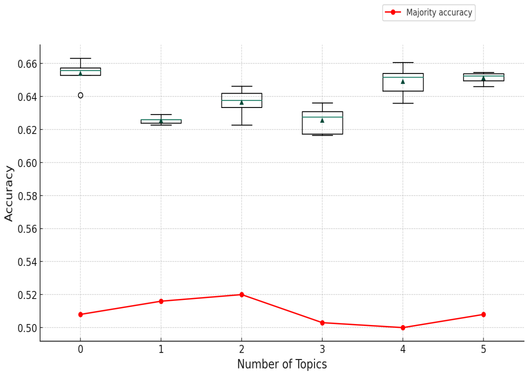


Figure 6: The performance of POLOR on unseen topics. The model shows a relatively stable performance and always outperforms the majority classifier even when we increase the number of hidden topics.

Cross-Topic Analysis. The baseline results in the quantitative section demonstrate that the model learns the political ideology of the news, not the source writing style since the test set was collected from unseen sources. However, that experiment falls short of explaining whether the model learns other textual properties such as topics. In this experiment, we are interested in measuring the model's generalizability on novel topics where we evaluate the model performance on unseen topics during testing. We iteratively increase the number of unseen topics (from 0 to 5 topics). We report the accuracy values across five different runs to account for the variations and uncontrollable randomness. For comparability reasons, we also include the majority classifier accuracy on the given split. We perform this experiment on the Large Media dataset since it consists of multiple topics. Figure 6 shows the results of this experiment. The model shows relatively stable performance even when we increase the number of hidden topics. The model always outperforms the majority classifier even when we increase the number of hidden topics. That is, the model is still capable of predicting article-specific labels on novel topics that were not seen during training.

	Majority	Acc	F1
Source annotations	0.74	0.91	0.91
Crowd-sourcing	0.51	0.68	0.66

Table 6: Comparing article-specific orientation with source-specific orientation, using the same test data but annotated in two different ways.

Benchmarking Task Difficulty. Extrapolate article-specific orientation from source annotations is less straightforward than predicting source-specific orientation. In this section, we benchmark the two tasks. For both tasks, the training data is propagated from source annotations. However, for the article-specific task, the test data is annotated via crowd-sourcing, while the test data for the source-specific task continues to be annotated via source annotations. Tabel 6 shows how predicting source-specific orientations of articles yields good performance because the model learns to map the predictions to sources rather than predicting the article orientation. On the other hand, the model struggles with learning article-specific orientation because the annotations in the test data are partially independent of the source annotations in the training data.

6 Conclusion

We propose POLOR, a fined-tuned BERT model augmented with multiple contrastive learning objectives to deduce article-specific labels even with training data annotated based on sources. Our model shows improved performance in predicting political orientation from unseen news sources by leveraging the similarities and differences between sentences. More specifically, the model tunes sentence embeddings by contrasting them to similar/different sentences. We introduce two datasets that span several U.S. news sources to evaluate the model performance. Moreover, our annotation results exhibit the relationship between sources and opinion articles by showing that opinion news pieces are usually driven by the author's beliefs, not the source. In future work, we attempt to explore the problem by incorporating additional knowledge other than text such as the author's background, personal experiences, and social interactions.

7 Limitations

Detecting political orientation can be helpful in various settings. However, dealing with the problem as a binary classification task can oversimplify the political landscape and fail to consider the wide cultural and regional variations. For example, a news article might hold conservative views on economic issues while being progressive on social issues. In that case, it is crucial to approach the problem as a multi-class classification problem where labels can be replaced with probabilities to portray different political views. Moreover, using crowd-sourcing platforms may not always yield the desired outcomes when annotating a text's political orientation. Workers should acquire a decent knowledge of the beliefs, perspectives, and variations of the political spectrum. Filtering mechanisms and increasing the number of annotators per task can partially solve the problem but such services can be quite expensive (twice the original price per task). Our attempt to develop a model incorporating different objective functions can yield decent results. However, choosing the appropriate objective functions can take a considerable amount of time. That is, incorporating a new objective function would always require performing a new parameters search since the weights (α and β) may drastically change depending on the incorporated objectives and the dataset itself.

References

[Bahdanau, Cho, and Bengio 2014] Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv e-prints* arXiv:1409.0473.

[Baly et al. 2020] Baly, R.; Da San Martino, G.; Glass, J.; and Nakov, P. 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)*, 4982–4991. Online: Association for Computational Linguistics.

[Budak, Goel, and Rao 2016] Budak, C.; Goel, S.; and Rao, J. M. 2016. Fair and Balanced? Quantifying Media Bias through Crowdsourced Content Analysis. *Public Opinion Quarterly* 80(S1):250–271.

[Card et al. 2015] Card, D.; Boydstun, A. E.; Gross, J. H.; Resnik, P.; and Smith, N. A. 2015. The media frames corpus: Annotations of frames across issues. In *Annual Meeting of the Association for Computational Linguistics*.

[Chen et al. 2018] Chen, W.-F.; Wachsmuth, H.; Al-Khatib, K.; and Stein, B. 2018. Learning to Flip the Bias of News Headlines. In *Proceedings of the 11th International Conference on Natural Language Generation*, 79–88. Tilburg University, The Netherlands: Association for Computational Linguistics.

[Chen et al. 2020] Chen, W.-F.; Al Khatib, K.; Wachsmuth, H.; and Stein, B. 2020. Analyzing Political Bias and Unfairness in News Articles at Different Levels of Granularity. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, 149–154. Online: Association for Computational Linguistics.

[de Jong 2023] de Jong, B. 2023. Why journalists are speaking out against western media bias in reporting on israel-palestine. Illustration by Walker Gawande, Edited by Tina Lee.

[Fan et al. 2019] Fan, L.; White, M.; Sharma, E.; Su, R.; Choubey, P. K.; Huang, R.; and Wang, L. 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)*, 6343–6349. Hong Kong, China: Association for Computational Linguistics.

[Gangula, Duggenpudi, and Mamidi 2019] Gangula, R. R. R.; Duggenpudi, S. R.; and Mamidi, R. 2019. Detecting political bias in news articles using headline attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 77–84. Florence, Italy: Association for Computational Linguistics.

[Gentzkow and Shapiro 2006] Gentzkow, M., and Shapiro, J. 2006. What Drives Media Slant? Evidence from U.S. Daily Newspapers. *Econometrica* 78:35–71.

[Greene and Resnik 2009] Greene, S., and Resnik, P. 2009. More than Words: Syntactic Packaging and Implicit Sentiment. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 503–511. Boulder, Colorado: Association for Computational Linguistics.

[Groseclose and Milyo 2005] Groseclose, T., and Milyo, J. 2005. A Measure of Media Bias. *The Quarterly Journal of Economics* 120(4):1191–1237.

[Jiang et al. 2019] Jiang, Y.; Petrak, J.; Song, X.; Bontcheva, K.; and Maynard, D. 2019. Team bertha von suttner at SemEval-2019 task 4: Hyperpartisan news detection using ELMo sentence representation convolutional network. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, 840–844. Minneapolis, Minnesota, USA: Association for Computational Linguistics.

[Kiesel et al. 2019] Kiesel, J.; Mestre, M.; Shukla, R.; Vincent, E.; Adineh, P.; Corney, D.; Stein, B.; and Potthast, M. 2019. SemEval-2019 task 4: Hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, 829–839. Minneapolis, Minnesota, USA: Association for Computational Linguistics.

[Kim and Johnson 2022] Kim, M. Y., and Johnson, K. M. 2022. CLoSE: Contrastive Learning of Subframe Embeddings for Political Bias Classification of News Media. In *Proceedings of the 29th International Conference on Computational Linguistics*, 2780–2793. Gyeongju, Republic of Korea: International Committee on Computational Linguistics.

[Kulkarni et al. 2018] Kulkarni, V.; Ye, J.; Skiena, S.; and Wang, W. Y. 2018. Multi-view Models for Political Ideology Detection of News Articles. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 3518–3527. Brussels, Belgium: Association for Computational Linguistics.

[Lazaridou et al. 2020] Lazaridou, K.; Löser, A.; Mestre, M.; and Naumann, F. 2020. Discovering biased news articles leveraging multiple human annotations. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 1268–1277. Marseille, France: European Language Resources Association.

[Lee et al. 2022] Lee, N.; Bang, Y.; Yu, T.; Madotto, A.; and Fung, P. 2022. Neus: Neutral multi-news summarization for mitigating framing bias.

[Liu et al. 2022] Liu, Y.; Zhang, X. F.; Wegsman, D.; Beauchamp, N.; and Wang, L. 2022. POLITICS: Pretraining with same-story article comparison for ideology prediction and stance detection. In *Findings of the Association for Computational Linguistics: NAACL 2022*, 1354–1374. Seattle, United States: Association for Computational Linguistics.

[Lundberg and Lee 2017] Lundberg, S. M., and Lee, S.-I. 2017. A unified approach to interpreting model predictions. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in*

Neural Information Processing Systems, volume 30. Curran Associates, Inc.

[Reimers and Gurevych 2019] Reimers, N., and Gurevych, I. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

[Roy and Goldwasser 2020] Roy, S., and Goldwasser, D. 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)*, 7698–7716. Online: Association for Computational Linguistics.

[Schroff, Kalenichenko, and Philbin 2015] Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 815–823.

[van der Maaten and Hinton 2008] van der Maaten, L., and Hinton, G. 2008. Viualizing data using t-sne. *Journal of Machine Learning Research* 9:2579–2605.

A High-Profile Cases

The articles were directly crawled from The Boston Globe, The Philadelphia Tribune, and The Wall Street Journal. We searched by keyword using the case name (i.e. Kyle Rittenhouse or Ahmaud Arbery). The articles are then manually inspected by the authors to ensure the articles specifically discuss the cases in the datasets. We have chosen the two controversial legal cases to create an annotated dataset because such cases polarized the news media and the articles they published strongly. Since the news articles were directly collected from the news outlet websites that are free to all readers, we do not expect to find harmful content. The following description of the two cases quoted directly from Wikipedia:

> *"**Ahmaud Arbery**, a 25-year-old black man, was murdered during a racially motivated hate crime while jogging in Satilla Shores, a neighborhood near Brunswick in Glynn County, Georgia. Erroneously assuming he was a burglar, three white men pursued Arbery in their trucks for several minutes, using the vehicles to block his path as he tried to run away. Two of the men, Travis McMichael and his father, Gregory McMichael, were armed in one vehicle. Their neighbor, William "Roddie" Bryan, was in another vehicle. After overtaking Arbery, Travis McMichael exited his truck and assaulted Arbery with a shotgun. As Arbery attempted to defend himself, Travis McMichael fatally shot him."*
>
> *"**Kyle Rittenhouse** (born January 3, 2003) is an American conservative activist who shot three men, two fatally, during the civil unrest in Kenosha, Wisconsin, in August 2020 when he was 17 years old. At his trial in November 2021, a jury found Rittenhouse not guilty of murder and all other charges after he testified that he acted in self-defense."*

B Dataset Construction

B.1 Source-Based Dataset

The dataset consists of 7 different news sources and addresses the two high-profile cases mentioned above. The text was pre-processed and cleaned using two Python packages, namely NLTK and Regular Expression(re). The sentences were further partitioned into sentences to feed them to the model. We provide detailed statistics across the news sources and the aforementioned cases in Table 7. We obtain the annotations from `AllSides.com` website. `AllSides.com` provides four different annotations obtained based on their rating methods: "Left", "Lean Left", "Lean Right", or "Right". For our study, we reduce the number of classes to two where "Left" and "Lean Left" are considered **Liberal**, and "Lean Right" and "Right" are considered **Conservative**. We follow the same annotation process of recently published datasets where the article annotations are propagated from their source annotations.

B.2 Human-Annotated Dataset

The crowd workers were provided with complete instructions about the cases ahead of time. Figure 7 shows an example paragraph and the instructions provided from the workers' point of view. We also made sure the workers had enough time to read the instructions and the paragraph carefully by setting the task timeout to 30 minutes. We also omit the source of each paragraph from workers to ensure their annotations are only based on the content. The paragraphs were examined and partitioned carefully such that they convey meaningful moments about the two cases. Table 8 shows the full statistics across the two topics and the news sources. Figure 8 shows a sample response from Amazon Mechanical Turk of an example paragraph.

For the final annotations, Amazon SageMaker Ground Truth results produce a confidence value for the label assigned to each paragraph [6]. If the annotators are all in agreement, then the confidence is high and low otherwise. Figure 9 show the distribution of the confidence levels for the 313 paragraphs the annotators have labeled. In the Appendix section, we show the label produced for an example paragraph. Among the 313 paragraphs, only 10 paragraphs received all three labels (i.e. the highest disagreement). This suggests the inter-annotator agreement was very high. For these 10 paragraphs, we chose the "Neutral" label pessimistically. We restricted each annotator to one paragraph to minimize any subjective bias introduced to the dataset. The annotators are U.S. residents and native speakers of English. We paid 0.25 dollars for each annotation(a total of 0.75 dollars per paragraph). Since we had selected only one annotator for each paragraph, we assumed no human bias. Below, we show the instructions given to the annotators:

"We are developing an algorithm to differentiate potential racial bias in news articles. You will be reading pieces of

[6] https://docs.aws.amazon.com/sagemaker/latest/dg/sms-data-output.html#sms-output-confidence

	CNN	WP	NYT	Chicago Tribune	LA	Fox	Houston Chronicle	
Arbery	20	20	20	19	17	21	13	130
Rittenhouse	20	20	19	20	20	14	20	133
Articles#	40	40	39	39	37	35	33	263
Sentences#	2365	1897	2213	1916	1434	923	1173	11921

Table 7: Detailed statistics of the Source-Based dataset. For each news source, we collected articles about the two cases and we ensured that the number of articles (*Articles#*) for both cases was approximately the same. Each article is then processed into sentences (*Sentences#*) to be fed to the model.

	BG	PT	WSJ	
Arbery	10	5	3	18
Rittenhouse	11	5	5	21
Articles#	21	10	8	39
Paragraphs#	170	78	65	313
Sentences#	623	353	247	1223

(a) Detailed statistics of the dataset.

	BG	PT	WSJ	
Liberal	83	35	20	138
Conservative	67	31	35	133
Neutral	20	12	10	42
Total#	170	78	65	313

(b) The crowd-sourcing annotations.

Table 8: An overview of the Human-Annotated dataset. The dataset spans the following new sources: BG (Boston Globe), PT (Philadelphia Tribune), and WSJ (Wall Street Journal).

texts (4-5 sentences long) extracted from real news media and you will be asked to determine the political ideology of the text whether Liberal or Conservative. If you are 100% unsure of the answer, choose Neutral. The text that you will be reading is mainly about two cases:

1. *Ahmaud Arbery: a black man who was shot and killed by three white men (Travis McMichael, Gregory McMichael, and, William Bryan) who said they were trying to protect their neighborhood from break-ins. [NY Times]*
2. *Kyle Rittenhouse: a 17-year-old who fatally shot two men and wounded another man in Kenosha, Wisconsin. The shootings occurred during the protests, riots, and civil unrest that followed the shooting of a black man, Jacob Blake, by a white police officer. Rittenhouse and those he shot were all white. [Wikipedia]*

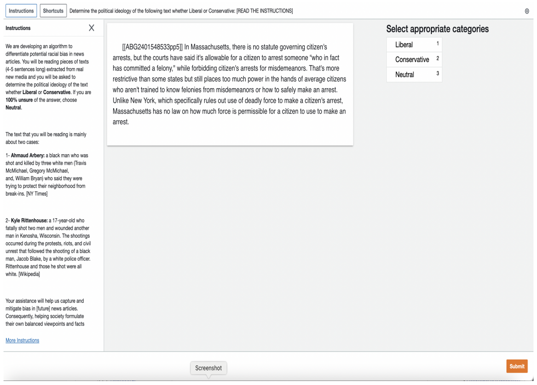
The following are instructions to be followed:

1. *You should be able to differentiate between Liberal and Conservative ideologies and their points of view on different aspects.*
2. *Read the instructions carefully and understand what you are expected to read.*
3. *Choose the appropriate labels that best suit the text."*

C Large Media Experiment

C.1 Dataset Collection and Annotation

Collection and Preprocessing. For this experiment, we used a dataset of news articles that spans 15 news sources from Budak, Goel, and Rao (2016). The news articles were collected from online news media websites published in 2013. Since the dataset was collected from online news websites that are free to all readers, we do not anticipate finding harmful content. The author first collected around 340,000 articles. Then, two binary classifiers were applied to filter



Determine the political ideology of the following text whether Liberal or Conservative: [READ THE INSTRUCTIONS]

Instructions

We are developing an algorithm to differentiate potential racial bias in news articles. You will be reading pieces of texts (4-5 sentences long) extracted from real new media and you will be asked to determine the political ideology of the text whether **Liberal** or **Conservative**. If you are **100% unsure** of the answer, choose **Neutral**.

The text that you will be reading is mainly about two cases:

1- **Ahmaud Arbery:** a black man who was shot and killed by three white men (Travis McMichael, Gregory McMichael, and, William Bryan) who said they were trying to protect their neighborhood from break-ins. [NY Times]

2- **Kyle Rittenhouse:** a 17-year-old who fatally shot two men and wounded another man in Kenosha, Wisconsin. The shootings occurred during the protests, riots, and civil unrest that followed the shooting of a black man, Jacob Blake, by a white police officer. Rittenhouse and those he shot were all white. [Wikipedia]

Your assistance will help us capture and mitigate bias in [future] news articles. Consequently, helping society formulate their own balanced viewpoints and facts

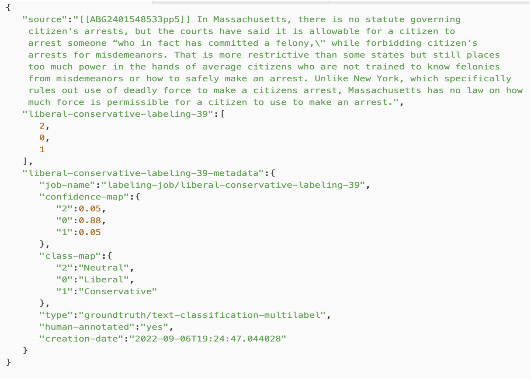
More Instructions

[[ABG2401548533pp5]] In Massachusetts, there is no statute governing citizen's arrests, but the courts have said it's allowable for a citizen to arrest someone "who in fact has committed a felony," while forbidding citizen's arrests for misdemeanors. That's more restrictive than some states but still places too much power in the hands of average citizens who aren't trained to know felonies from misdemeanors or how to safely make an arrest. Unlike New York, which specifically rules out use of deadly force to make a citizen's arrest, Massachusetts has no law on how much force is permissible for a citizen to use to make an arrest.

Select appropriate categories

Liberal	1
Conservative	2
Neutral	3

Figure 7: An example of a paragraph annotation task that shows the complete instructions provided to workers.



```
{
   "source":"[[ABG2401548533pp5]] In Massachusetts, there is no statute governing
    citizen's arrests, but the courts have said it is allowable for a citizen to
    arrest someone “who in fact has committed a felony,\" while forbidding citizen's
    arrests for misdemeanors. That is more restrictive than some states but still places
    too much power in the hands of average citizens who are not trained to know felonies
    from misdemeanors or how to safely make an arrest. Unlike New York, which specifically
    rules out use of deadly force to make a citizens arrest, Massachusetts has no law on how
    much force is permissible for a citizen to use to make an arrest.",
   "liberal-conservative-labeling-39":[
      2,
      0,
      1
   ],
   "liberal-conservative-labeling-39-metadata":{
      "job-name":"labeling-job/liberal-conservative-labeling-39",
      "confidence-map":{
         "2":0.05,
         "0":0.88,
         "1":0.05
      },
      "class-map":{
         "2":"Neutral",
         "0":"Liberal",
         "1":"Conservative"
      },
      "type":"groundtruth/text-classification-multilabel",
      "human-annotated":"yes",
      "creation-date":"2022-09-06T19:24:47.044028"
   }
}
```

Figure 8: A sample response from Amazon Mechanical Turk of an annotation task.

out political news. This resulted in 115,000 political news articles. The author then used Amazon Mechanical Turk to an-

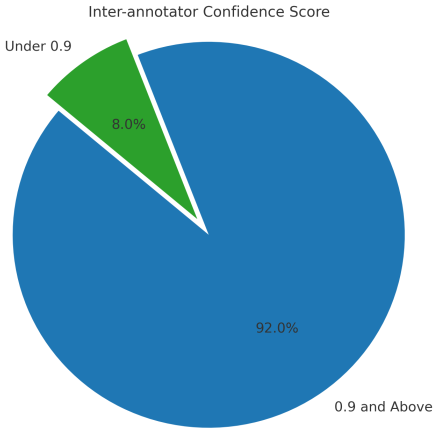


Figure 9: The distribution of inter-annotator agreement score among the three annotators per paragraph.

Topic	Count
Elections	924
Healthcare	882
Economy	841
None	740
Democrat Scandals	511
International News	395
National Security	377
Civil Rights	377
Republican Scandals	365
Gun Rights Regulation	338
Gay Rights	135
Gun Related Crimes	66
Environment	62
Education	59
Drugs	17
Total	6089

Table 9: The distribution of Topics in the large media dataset.

notate the article’s topic and orientation. The workers were asked to pick a primary and secondary topic out of 15 possible topics. Then, the workers were asked to annotate the political orientation scale of the Democratic and Republican parties. The scale is encoded as ”positive”, ”somewhat positive”, ”neutral”, ”somewhat negative”, and ”negative”. The dataset is freely available at the University of Michigan - Deep Blue Data website [7]. The published dataset contains only 21,000 articles. Each article has the following attributes:

1. Url
2. News Type: other, News, or Opinion.
3. Perceived: whether the worker was looking at the blinded or unblinded version.
4. Primary topic
5. Secondary topic
6. Democratic party vote
7. Republican vote

The actual articles (i.e. the body of articles) were not provided in the dataset and only the URLs to these articles were published. To collect the text of the articles, we used the `newspaper` package in Python. We could not retrieve all articles from the provided links due to several reasons, such as URL outdated, HTML format change, and package failure to process articles. We were able to retrieve a subset of 15,780 articles. The final retrieved articles are further processed by eliminating unnecessary information such as identifying content, punctuation, non-English characters, Twitter mentions, and URLs.

Annotation. Since we are interested in one annotation per article (i.e. Liberal, Conservative, or Neutral), we synthesize a universal annotation by setting some rules on the columns ”Democratic party vote” and ”Republican vote”. Since both of them have the same scale, we assign the label ”Democrat” if the Democrat vote score is higher than the Republican vote score. On the other hand, the ”Republican” label is assigned if the Republican vote score is higher. We assign ”Neutral” in all other cases. For the source annotations, the remaining articles were annotated based on the reviews of each source found on `AllSides.com` website. Table 10 shows the orientation of sources (publishers) as they appear in `AllSides.com`. Then, the articles are filtered to ensure that both the source label and human label belong to $\{Liberal, Conservative\}$. That is, articles labeled as neutral based on their source orientations or crowd-sourcing annotations were removed. The final number of articles after filtering label neutral labels is 6089 articles.

The final dataset consists of 15 topics as shown in Table 9. The ”None” topic is removed during the Zero-shot learning experiment in the Cross-Media experiment. The dataset shows a huge class imbalance when it’s labeled by sources. This class imbalance is automatically solved when we divide the dataset into training, validation, and testing splits. That is, the training data is labeled by source, but the testing and validation are both labeled by the crowd-sourcing labels (human labels). We split the dataset for all experiments as 70% for training, 15% for validation, and 15% for testing. We first divide the dataset into 70% and 30% splits by stratifying the source label to fix the class imbalance. The validation and testing are then derived from the 30% split. The distribution of the number of articles and labels in each split is shown in Table 11.

C.2 Experiment Configurations

We trained the model on articles for all experiments that use this dataset. Since the BERT model only takes input sequences up to 512 tokens, the first 512 tokens of each article were used as our input. The reason behind using articles, not sentences is to reduce the runtime. The best-performing parameters in the baseline experiment (Section 5.3) were used

[7] https://deepblue.lib.umich.edu/data/concern/data_sets/8w32r569d

Publisher	Orientation	Count
Breitbart News	right	736
CNN	left	319
Daily Kos	left	906
Fox News	right	645
Huffington Post	left	541
Los Angeles Times	left	458
NBC News	left	498
New York Times	left	558
USA Today	left	496
Washington Post	left	671
Yahoo News	left	261
Total		6089

Table 10: The distribution of publishers in the large media dataset.

Split	Liberal	Conservative	Total
Train	3296	967	4263
Validation	476	437	913
Test	476	437	913
Total	-	-	6089

Table 11: The distribution of labels across splits in the large media study. The labels in the training data are source annotated, while the testing and validation are human annotated.

throughout all experiments when applicable. For example, we used $\alpha = 0$ and $\beta = 0$ to predict topics, publishers, and source labels in Table 2. For the Cross-Media experiment in Section 5.4, the results were averaged across five different for each data point in Figure 6. The topics and the distribution of articles used in the experiment can be found in Table 9.