We Can Detect Your Bias: Predicting the Political Ideology of News Articles

Ramy Baly[1], **Giovanni Da San Martino**[2], **James Glass**[1], **Preslav Nakov**[2]
[1]MIT Computer Science and Artificial Intelligence Laboratory
[3]Qatar Computing Research Institute, HBKU
{baly,glass}@mit.edu
{gmartino,pnakov}@hbku.edu.qa

Abstract

We explore the task of predicting the leading political ideology or bias of news articles. First, we collect and release a large dataset of 34,737 articles that were manually annotated for political ideology –left, center, or right–, which is well-balanced across both topics and media. We further use a challenging experimental setup where the test examples come from media that were not seen during training, which prevents the model from learning to detect the source of the target news article instead of predicting its political ideology. From a modeling perspective, we propose an adversarial media adaptation, as well as a specially adapted triplet loss. We further add background information about the source, and we show that it is quite helpful for improving article-level prediction. Our experimental results show very sizable improvements over using state-of-the-art pre-trained Transformers in this challenging setup.

1 Introduction

In any piece of news, there is a chance that the viewpoint of its authors and of the media organization they work for, would be reflected in the way the story is being told. The emergence of the Web and of social media has lead to the proliferation of information sources, whose leading political ideology or bias may not be explicit. Yet, systematic exposure to such bias may foster intolerance as well as ideological segregation, and ultimately it could affect voting behavior, depending on the degree and the direction of the media bias, and on the voters' reliance on such media (DellaVigna and Kaplan, 2007; Iyengar and Hahn, 2009; Saez-Trumper et al., 2013; Graber and Dunaway, 2017). Thus, making the general public aware, e.g., by tracking and exposing bias in the news is important for a healthy public debate given the important role media play in a democratic society.

Media bias can come in many different forms, e.g., by omission, by over-reporting on a topic, by cherry-picking the facts, or by using propaganda techniques such as appealing to emotions, prejudices, fears, etc. (Da San Martino et al., 2019, 2020a,b) Bias can occur with respect to a specific topic, e.g., COVID-19, immigration, climate change, gun control, etc. (Darwish et al., 2020; Stefanov et al., 2020) It could also be more systematic, as part of a political ideology, which in the Western political system is typically defined as left vs. center vs. right political leaning.

Predicting the bias of individual news articles can be useful in a number of scenarios. For news media, it could be an important element of internal quality assurance as well as of internal or external monitoring for regulatory compliance. For news aggregator applications, such as Google News, it could enable balanced search, similarly to what is found on AllSides.[1] For journalists, it could enable news exploration from a left/center/right angle. It could also be an important building block in a system that detects bias at the level of entire news media (Baly et al., 2018, 2019, 2020), such as the need to offer explainability, i.e., if a website is classified as left-leaning, the system should be able to pinpoint specific articles that support this decision.

In this paper, we focus on predicting the bias of news articles as left-, center-, or right-leaning. Previous work has focused on doing so at the level of news media (Baly et al., 2020) or social media users (Darwish et al., 2020), but rarely at the article level (Kulkarni et al., 2018). The scarce article-level research has typically used distant supervision, assuming that all articles from a given medium should share its overall bias, which is not always the case. Here, we revisit this assumption.

[1]http://allsides.com/

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 4982–4991,
November 16–20, 2020.

Our contributions can be summarized as follows:

- We create a new dataset for predicting the political ideology of news articles. The dataset is annotated at the article level and covers a wide variety of topics, providing balanced left/center/right perspectives for each topic.
- We develop a framework that discourages the learning algorithm from modeling the source instead of focusing on detecting bias in the article. We validate this framework in an experimental setup where the test articles come from media that were not seen at training time. We show that adversarial media adaptation is quite helpful in that respect, and we further propose to use a triplet loss, which shows sizable improvements over state-of-the-art pre-trained Transformers.
- We further incorporate media-level representation to provide background information about the source, and we show that this information is quite helpful for improving the article-level prediction even further.

The rest of this paper is organized as follows: We discuss related work in Section 2. Then, we introduce our dataset in Section 3, we describe our models for predicting the political ideology of a news article in Section 4, and we present our experiments and we discuss the results in Section 5. Finally, we conclude with possible directions for future work in Section 6.

2 Related Work

Most existing datasets for predicting the political ideology at the news article level were created by crawling the RSS feeds of news websites with known political bias (Kulkarni et al., 2018), and then projecting the bias label from a website to all articles crawled from it, which is a form of distant supervision. The crawling could be also done using text search APIs rather than RSS feeds (Horne et al., 2019; Gruppi et al., 2020).

The media-level annotation of political leaning is typically obtained from specialized online platforms, such as News Guard,[2] AllSides,[3] and Media Bias/Fact Check,[4] where highly qualified journalists use carefully designed guidelines to make the judgments.

[2] http://www.newsguardtech.com
[3] http://allsides.com/
[4] http://mediabiasfactcheck.com

As manual annotation at the article level is very time-consuming, requires domain expertise, and it could be also subjective, such annotations are rarely available at the article level. As a result, automating systems for political bias detection have opted for using distant supervision as an easy way to obtain large datasets, which are needed to train contemporary deep learning models.

Distant supervision is a popular technique for annotating datasets for related text classification tasks, such as detecting hyper-partisanship (Horne et al., 2018; Potthast et al., 2018) and propaganda/satire/hoaxes (Rashkin et al., 2017). For example, Kiesel et al. (2019) created a large corpus for detecting hyper-partisanship (i.e., articles with extreme left/right bias) consisting of 754,000 articles, annotated via distant supervision, and additional 1,273 manually annotated articles, part of which was used as a test set for the SemEval-2019 task 4 on Hyper-partisan News Detection. The winning system was an ensemble of character-level CNNs (Jiang et al., 2019). Interestingly, all top-performing systems in the task achieved their best results when training on the manually annotated articles only and ignoring the articles that were labeled using distant supervision, which illustrates the dangers of relying on distant supervision.

Barrón-Cedeno et al. (2019) extensively discussed the limitations of distant supervision in a text classification task about article-level propaganda detection, in a setup that is similar to what we deal with in this paper: the learning systems may learn to model the source of the article instead of solving the task they are actually trained for. Indeed, they have shown that the error rate may drastically increase if such systems are tested on articles from sources that were never seen during training, and that this effect is positively correlated with the representation power of the learning model. They analyzed a number of representations and machine learning models, showing which ones tend to overfit more, but, unlike our work here, they fell short of recommending a practical solution.

Budak et al. (2016) measured the bias at the article level using crowd-sourcing. This is risky as public awareness of media bias is limited (Elejalde et al., 2018). Moreover, the annotation setup does not scale. Finally, their dataset is not freely available, and their approach of randomly crawling articles does not ensure that topics and events are covered from different political perspectives.

Lin et al. (2006) built a dataset annotated with the ideology of 594 articles related to the Israeli-Palestinian conflict published on bitterlemons.org. The articles were written by two editors and 200 guests, which minimizes the risk of modeling the author style. However, the dataset is too small to train modern deep learning approaches.

Kulkarni et al. (2018) built a dataset using distant supervision and labels from AllSides. Distant supervision is fine for the purpose of training, but they also used it for testing, which can be problematic. Moreover, their training and test sets contain articles from the same media, and thus models could easily learn to predict the article's source rather than its bias. In their models, they used both the text and the URL contents of the articles.

Overall, political bias has been studied at the level of news outlet (Dinkov et al., 2019; Baly et al., 2018, 2020; Zhang et al., 2019), user (Darwish et al., 2020), article (Potthast et al., 2018; Saleh et al., 2019), and sentence (Sim et al., 2013; Saez-Trumper et al., 2013). In particular, Baly et al. (2018) developed a system to predict the political bias and the factuality of news media. In a follow-up work, Baly et al. (2019) showed that bias and factuality of reporting should be predicted jointly. A finer-grained analysis is performed in (Horne et al., 2018), where a model was trained on 10K sentences from a dataset of reviews (Pang and Lee, 2004), and used to discriminate objective versus non-objective sentences in news articles. Lin et al. (2006) presented a sentence-level classifier, where the labels were projected from the document level.

3 Dataset

In this section, we describe the dataset that we created and that we used in our experiments. While most of the platforms that analyze the political leaning of news media provide in-depth analysis of particular aspects of the media, AllSides stands out as it provides annotations of political ideology for individual articles, which ensures high-quality data for both training and testing, which is in contrast with distant supervision approaches used in most previous research, as we have seen above. In AllSides, these annotations are made as a result of a rigorous process that involves blind bias surveys, editorial reviews, third-party analysis, independent reviews, and community feedback.[5]

Furthermore, AllSides uses the annotated articles to enable its *Balanced Search*, which shows news coverage on a given topic from media with different political bias. In other words, for each trending event or topic (e.g., *impeachment* or *coronavirus pandemic*), the platform pushes news articles from all sides of the political spectrum, as shown in Figure 1. We took advantage of this and downloaded all articles along with their political ideology annotations (*left*, *center*, or *right*), their assigned topic(s), the media in which they were published, their author(s), and their publication date. Thus, our dataset contains articles that were manually selected and annotated, and that are representative of the real political scenery. Note that the *center* class covers articles that are biased towards a centrist political ideology, and not articles that lack political bias (e.g., *sports* and *technology*), which commonly exist in news corpora that were built by scraping RSS feeds.

We collected a total of 34,737 articles published by 73 news media and covering 109 topics.[6] In this dataset, a total of 1,080 individual articles (3.11%) have a political ideology label that is different from their source's. This suggests that, while the distant supervision assumption generally holds, we would still find many articles that defy it. Table 1 shows some statistics about the dataset.

Political Ideology	Count	Percentage
Left	12,003	34.6%
Center	9,743	28.1%
Right	12,991	37.3%

Table 1: Statistics about our dataset.

Figure 2 illustrates the distribution of the different political bias labels within each of the most frequent topics. We can see that our dataset is able to represent topics or events from different political perspectives. This is yet another advantage, as it enables a more challenging task for machine learning models to detect the linguistic and the semantic nuances of different political ideologies in news articles, as opposed to cases where certain topics might be coincidentally collocated with certain labels, in which case the models would be actually learning to detect the topics instead of predicting the political ideology of the target news article.

[5] http://www.allsides.com/media-bias/media-bias-rating-methods

[6] In some cases, an article could be assigned to multiple topics, e.g., it could go simultaneously into *coronavirus*, *public health*, and *healthcare*.

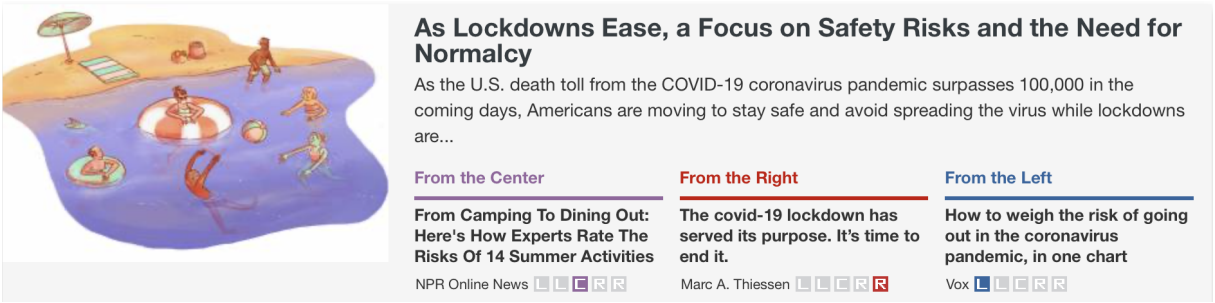

Figure 1: AllSides: balanced search on the topic of *reopening after the coronavirus lockdown*.

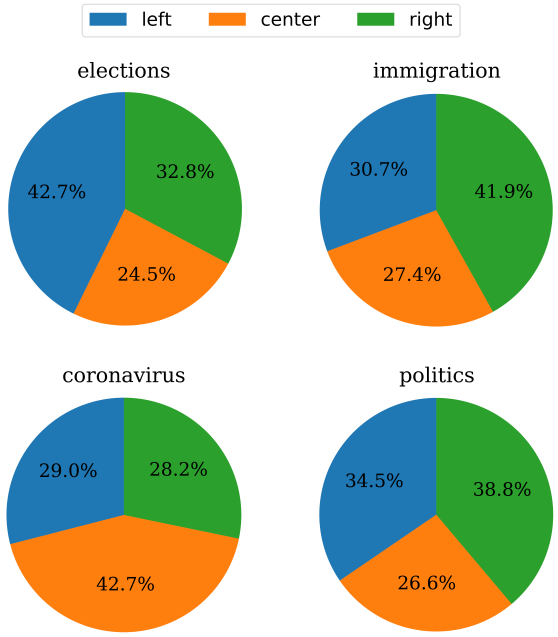

Figure 2: Political ideology for the most frequent topics: *elections*, *immigration*, *coronavirus*, and *politics*.

It is worth noting that since most article labels are aligned with their source labels, it is likely that machine learning classifiers would end up modeling the source instead of the political ideology of the individual articles. For example, a model would be learning the writing style of each medium, and then it would associate it with a particular ideology. Therefore, we pre-processed the articles in a way that eliminates explicit markers such as the name of the authors, or the name of the medium that usually appears as a preamble to the article's content, or in the content itself. Furthermore, in order to ensure that we are actually modeling the political ideology as it is expressed in the language of the news, we created evaluation splits in two different ways: (*i*) randomly, which is what is typically done (for comparison only), and (*ii*) based on media, where all articles by the same medium appear in either the training, the validation, or the testing dataset.

The latter form of splitting would help us indicate what a trained classifier has actually learned. For instance, if it modeled the source, then it would not be able to perform well on the test set, since all its articles would belong to sources that were never seen during training. In order to ensure fair one-to-one comparisons between experiments, we created these two different sets of splits, while making sure that they share the same test set, as follows:

- **Media-based Split:** We sampled 1,200 articles from 12 news media (100 per medium) and used them as the *test* set, and we excluded the remaining 5,470 articles from these media. Then, we used the articles from the remaining 61 media to create the *training* and the *validation* sets, where all articles from the same medium would appear in the same set: training, development, or testing. This ensures that the model is fine-tuned and tested on articles whose sources were not seen during training.

- **Random Split:** Here, the *test* set is the same as in the media-based split. The 5,470 articles that we excluded from the 12 media are now added to the articles from the 61 remaining media. Then, we split this collection of articles (using stratified random sampling) into *training* and *validation* sets. This ensures that the model is fine-tuned and evaluated only on articles whose sources were observed during training.

Table 2 shows statistics about both splits, including the size of each set and the number of media and topics they cover. We release the dataset, along with the evaluation splits, and the code,[7] which can be used to extend the dataset as more news articles are added to AllSides.

[7] http://github.com/ramybaly/Article-Bias-Prediction

		Train	Valid.	Test
Media-based	*Count*	22,969	5,098	1,200
	Media	46	15	12
	Topics	108	105	93
Random	*Count*	26,828	6,709	1,200
	Media	73	73	12
	Topics	108	107	93

Table 2: Statistics about our dataset and its two splits: *media-based* and *random*.

4 Methodology

4.1 Classifiers

The task of predicting the political ideology of news articles is typically formulated as a classification problem, where the textual content of the articles is encoded into a vector representation that is used to train a classifier to predict one of C classes (in our case, $C = 3$: *left*, *center*, and *right*). In our experiments, we use two deep learning architectures: (*i*) *Long Short-Term Memory networks* (LSTMs), which are Recurrent Neural Networks (RNNs), which use gating mechanisms to selectively pass information across time and to model long-term dependencies (Hochreiter and Schmidhuber, 1997), and (*ii*) *Bidirectional Encoder Representations from Transformers* (BERT), with a complex architecture yielding high-quality contextualized embeddings, which have been successful in several Natural Language Processing tasks (Devlin et al., 2019).

4.2 Removing Media Bias

Ultimately, our goal is to develop a model that can predict the political ideology of a news article. Our dataset, along with some others, has a special property that might stand in the way of achieving this goal. Most articles published by a given source have the same ideological leaning. This might confuse the model and cause it to erroneously associate the output classes with features that characterize entire media outlets (such as detecting specific writing patterns, or stylistic markers in text). Consequently, the model would fail when applied to articles that were published in media that were unseen during training. The experiments in Section 5 confirm this. Thus, we apply two techniques to *de-bias* the models, i.e., to prevent them from learning the style of a specific news medium rather than predicting the political ideology of the target news article.

4.2.1 Adversarial Adaptation (AA)

This model was originally proposed by Ganin et al. (2016) for unsupervised domain adaptation in image classification. Their objective was to adapt a model trained on labelled images from a *source* domain to a novel *target* domain, where the images have no labels for the task at hand. This is done by adding an adversarial *domain classifier* with a gradient reversal layer to predict the examples' domains. The *label predictor's* is minimized for the labelled examples (from the source domain), and the adversarial *domain classifier's* loss is maximized for all examples in the dataset. As a result, the encoder can extract representation that is (*i*) discriminative for the main task and also (*ii*) invariant across domains (due to the gradient reversal layer). The overall loss is minimized as follows:

$$\sum_{\substack{i=1:N \\ d_i=0}} \mathcal{L}_y^i(\theta_f, \theta_y) - \lambda \sum_{i=1:N} \mathcal{L}_d^i(\theta_f, \theta_d), \qquad (1)$$

where N is the number of training examples, $\mathcal{L}_y^i(\cdot,\cdot)$ is the label predictor's loss, the condition $d_i = 0$ means that only examples from the source domain are used to calculate the label predictor's loss, $\mathcal{L}_d^i(\cdot,\cdot)$ is the domain classifier's loss, λ controls the trade-off between both losses, and $\{\theta_f, \theta_y, \theta_d\}$ are the parameters of the encoder, the label predictor, and the domain classifier, respectively. Further details about the formulation of this method is available in (Ganin et al., 2016).

We adapt this architecture as follows. Instead of a *domain classifier*, we implement a *media classifier*, which, given an article, tries to predict the medium it comes from. As a result, the encoder should extract representation that is discriminative for the main task of predicting political ideology, while being invariant for the different media. This approach was originally proposed as an unsupervised domain adaptation, since labelled examples were available for one domain only, whereas in our case, all articles from different media were labelled for their political ideology. Therefore, we jointly minimize the losses of both the *label predictor* and the *media classifier* over the entire dataset. The new objective function to minimize is as follows:

$$\sum_{i=1:N} \mathcal{L}_y^i(\theta_f, \theta_y) - \lambda \sum_{i=1:N} \mathcal{L}_m^i(\theta_f, \theta_m), \qquad (2)$$

where $\mathcal{L}_m^i(\cdot,\cdot)$ is the loss of the *media classifier*, and θ_m is its set of parameters.

4.2.2 Triplet Loss Pre-training (TLP)

In this approach, we pre-train the encoder using a triplet loss (Schroff et al., 2015). The model is trained on a set of triplets, each composed of an anchor, a positive, and a negative example. The objective in Eq. 3 ensures that the positive example is always closer to the anchor than the negative example is, where $\boldsymbol{a}$, $\boldsymbol{p}$ and $\boldsymbol{n}$ are the encodings of the anchor, of the positive, and of the negative examples, respectively, and $D(\cdot, \cdot)$ is the Euclidean distance:

$$\mathcal{L} = \max\left(D\left(\boldsymbol{a}, \boldsymbol{p}\right) - D\left(\boldsymbol{a}, \boldsymbol{n}\right) + \epsilon, 0\right). \quad (3)$$

Figure 3 shows an example of such a triplet. The positive example shares the same ideology as the anchor's, but they are published by different media. The negative example has a different ideology than the anchor's, but they are published by the same medium. In this way, the encoder will be clustering examples with similar ideologies close to each other, regardless of their source. Once the encoder has been pre-trained, its parameters, along with the softmax classifier's, are fine-tuned on the main task by minimizing the cross-entropy loss when predicting the political ideology of articles.

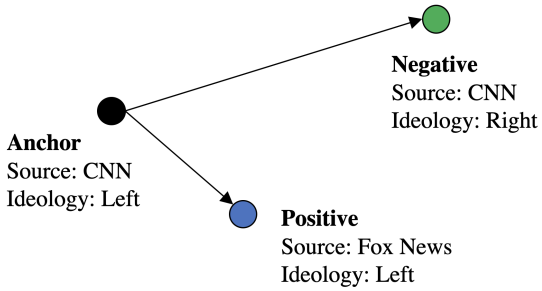

Figure 3: An example triplet used for de-biasing.

4.3 Media-level Representation

Finally, we explore the benefits of incorporating information describing the target medium, which can serve as a complementary representation for the article. While this seems to be counter-intuitive to what we have been proposing in Subsection 4.2, we believe that medium-level representation can be valuable when combined with an accurate representation of the article. Intuitively, having an accurate understanding of the natural language in the article, together with a glimpse into the medium it is published in, should provide a more complete picture of its underlying political ideology.

Baly et al. (2020) proposed a comprehensive set of representation to characterize news media from different angles: how a medium portrays itself, who is its audience, and what is written about it. Their results indicate that exploring the *Twitter bios* of a medium's followers offers a good insight into its political leaning. To a lesser extent, the content of a *Wikipedia* page describing a medium can also help unravel its political leaning. Therefore, we concatenated these representations to the encoded articles, at the output of the encoder and right before the SOFTMAX layer, so that both the article encoder and the classification layer that is based on the article and the external media representations are trained jointly and end-to-end.

Similarly to (Baly et al., 2020), we retrieved the profiles of up to a 1,000 Twitter followers for each medium, we encoded their bios using the Sentence-BERT model (Reimers and Gurevych, 2019), and we then averaged these encodings to obtain a single representation for that medium. As for the Wikipedia representation, we automatically retrieved the content of the page describing each medium, whenever applicable. Then, we used the pre-trained base BERT model to encode this content by averaging the word representations extracted from BERT's second-to-last layer, which is common practice, since the last layer may be biased towards the pre-training objectives of BERT.

5 Experiments and Results

We evaluated both the LSTM and the BERT models, assessing the impact of (*i*) de-biasing and (*ii*) incorporating media-level representation.

5.1 Experimental Setup

We fine-tuned the hyper-parameters of both models on the validation set using a guided grid search trial while fixing the seeds of the random weights initialization. For LSTM, we varied the length of the input (128–1,024 tokens), the number of layers (1–3), the size of the LSTM cell (200–400), the dropout rate (0–0.8), the learning rate ($1e{-}3$ to $1e{-}5$), the gradient clipping value (0–5), and the batch size (8–256). The best results were obtained with a 512-token input, a 2-layer LSTM of size 256, a dropout rate of 0.7, a learning rate of $1e{-}3$, gradient clipping at 0.5, and a batch size of 32. This model has around 1.1M trainable parameters, and was trained with 300-dimensional GloVe input word embeddings (Pennington et al., 2014).

For BERT, we varied the length of the input, the learning rate, and the gradient clipping value. The best results were obtained using a 512-token input, a learning rate of $2e{-}5$, and gradient clipping at 1. This model has 110M trainable parameters.

We trained our models on 4 *Titan X Pascal* GPUs, and the runtime for each epoch was 25 seconds for the LSTM-based models and 22 minutes for the BERT-based models. For each experiment, the model was trained only once with fixed seeds used to initialize the models' weights.

For the Adversarial Adaptation (AA), we have an additional hyper-parameter λ (see Equation 2), which we varied from 0 to 1, where 0 means no adaptation at all. The best results were obtained with $\lambda = 0.7$, which means that we need to pay significant attention to the adversarial classifier's loss in order to mitigate the media bias.

For the Triplet Loss Pre-training (PLT), we sampled 35,017 triplets from the training set, such that the examples in each triplet discuss the same topic in order to ensure that the change in topic has minimal impact on the distance between the examples.

To evaluate our models, we use accuracy and macro-F_1 score (F_1 averaged across all classes), which we also used as an early stopping criterion, since the classes were slightly imbalanced. Moreover, given the ordinal nature of the labels, we report the Mean Absolute Error (MAE), shown in Equation (4), where N is the number of instances, and y_i and $\hat{y}_i$ are the number of correct and of predicted labels, respectively.

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \quad (4)$$

5.2 Results

Baseline Results The results in Table 3 show the performance for LSTM and for BERT at predicting the political ideology of news articles for both the *media-based* and the *random* splits. We observe sizable differences in performance between the two splits. In particular, both models perform much better when they are trained and evaluated on the *random* split, whereas they both fail on the *media-based* split, where they are tested on articles from media that were not seen during training. This observation confirms our initial concerns that the models would tend to learn general characteristics about news media, and then would face difficulties with articles coming from new unseen media.

Model	Split	Macro F_1	Acc.	MAE
Majority		19.61	41.67	0.92
LSTM	*Media-based*	31.51	32.30	0.97
	Random	65.50	66.17	0.52
BERT	*Media-based*	35.53	36.75	0.90
	Random	80.19	79.83	0.33

Table 3: Baseline experiments (without de-biasing or media-level representation) for the two splits.

Removing the Source Bias In order to further confirm the bias towards modeling the media, we ran a side experiment of fine-tuning BERT on the task of predicting the medium given the article's content, which is a 73-way classification problem. We used stratified random sampling to create the evaluation splits and to make sure each set contains all labels (media). The results in Table 4 confirm that BERT is much stronger than the majority class baseline, despite the high number of classes, which means that predicting the medium in which a target news article was published is a fairly easy task.

Model	Macro F_1	Acc.
Majority	0.25	10.21
BERT	59.72	80.12

Table 4: Predicting the medium in which a target news article was published.

In order to remove the bias towards modeling the medium, we evaluated the impact of the adversarial adaptation (AA) and the Triplet Loss Pre-training (TLP) with the media-based split. The results in Table 5 show sizeable improvements when either of these approaches is used, compared to the baseline (no de-biasing). In particular, TLP yields an improvement of 14.12 points absolute in terms of accuracy, and 12.73 points in terms of macro-F_1.

Model	De-bias	Macro F_1	Acc.	MAE
LSTM	*None*	31.51	32.30	0.97
	AA	40.33	40.57	0.69
	TLP	**45.44**	**46.42**	**0.62**
BERT	*None*	35.53	36.75	0.90
	AA	43.87	46.22	0.59
	TLP	**48.26**	**51.41**	**0.51**

Table 5: Impact of de-biasing (adversarial adaptation and triplet loss) on article-level bias detection.

#	Representation	LSTM Macro F_1	LSTM Acc.	LSTM MAE	BERT Macro F_1	BERT Acc.	BERT MAE
1	*Article (baseline)*	31.51	32.30	0.97	35.53	36.75	0.90
2	*Article with TLP*	45.44	46.42	0.62	48.26	51.41	0.51
3	*Wikipedia*	41.39	41.86	0.92	41.39	41.86	0.92
4	*Wikipedia + Article*	40.49	40.79	0.92	42.33	41.90	0.90
5	*Wikipedia + Article with TLP*	48.25	46.47	0.69	51.16	49.75	0.32
6	*Twitter bios*	60.30	62.69	0.42	60.30	62.69	0.42
7	*Twitter bios + Article*	60.30	62.69	0.42	60.42	63.12	0.40
8	*Twitter bios + Article with TLP*	**62.02**	**70.03**	**0.32**	**64.29**	**72.00**	**0.29**

Table 6: Impact of adding media-level representations to the article-level representations (with and without de-biasing). Note that the results in rows 3 and 6 are the same for both LSTM and BERT because no articles were involved, and the media-level representations were directly used to train the classifier.

Impact of Media-Level Representation Finally, we evaluated the impact of incorporating the media-level representation (Twitter followers' bios and Wikipedia content) in addition to teh article-level representation. Table 6 illustrates these results in an incremental way. First, we evaluated the performance of the media-level representation alone at predicting the political ideology of news articles (see rows 3 and 6). We should note that these results are identical for the LSTM and the BERT columns since no article was encoded in these experiments, and the media representation was used directly to train the logistic regression classifier. Then, adding the article representation from either model, without any de-biasing, had no or little impact on the performance (see rows 4 vs. 3, and 7 vs. 6). This is not surprising, since we have shown that, without de-biasing, both models learn more about the source than about the bias in the language used by the article. Therefore, the ill-encoded articles do not provide more information than what the medium representation already gives, which is why no or too little improvement was observed.

When we use the triplet loss to mitigate the source bias, the resulting article representation is more accurate and meaningful, and the medium representation does offer complementary information, and eventually contributes to sizeable performance gains (see rows 5 and 8 vs. 2). The Twitter bios representation appears to be much more important than the representation from Wikipedia, which shows the importance of inspecting the media followers' background and their point of views, which is also one of the observations in (Baly et al., 2020).

Overall, comparing the best results to the baseline (rows 8 vs. 1), we can see that (*i*) using the triplet loss to remove the source bias, and (*ii*) incorporating media-level representation from Twitter followers yields 30.51 and 28.76 absolute improvement in terms of macro F_1 on the challenging *media-based* split.

6 Conclusion and Future Work

We have explored the task of predicting the leading political ideology of news articles. In particular, we created a new large dataset for this task, which features article-level annotations and is well-balanced across topics and media. We further proposed an adversarial media adaptation approach, as well as a special triplet loss in order to prevent modeling the source instead of the political bias in the news article, which is a common pitfall for approaches dealing with data that exhibit high correlation between the source of a news article and its class, as is the case with our task here. Finally, our experimental results have shown very sizable improvements over using state-of-the-art pre-trained Transformers.

In future work, we plan to explore topic-level bias prediction as well as going beyond left-center-right bias. We further want to develop models that would be able to detect specific fragments in an article where the bias occurs, thus enabling explainability. Last but not least, we plan to experiment with other languages, and to explore to what extent a model for one language is transferable to another one given that the left-center-right division is not universal and does not align perfectly across countries and cultures, even when staying within the Western political world.

Acknowledgments

This research is part of the Tanbih project[8], which aims to limit the effect of "fake news," propaganda and media bias by making users aware of what they are reading. The project is developed in collaboration between the Qatar Computing Research Institute, HBKU and the MIT Computer Science and Artificial Intelligence Laboratory.

References

Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018. Predicting factuality of reporting and bias of news media sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, EMNLP '18, pages 3528–3539, Brussels, Belgium.

Ramy Baly, Georgi Karadzhov, Jisun An, Haewoon Kwak, Yoan Dinkov, Ahmed Ali, James Glass, and Preslav Nakov. 2020. What was written vs. who read it: News media profiling using text analysis and social media context. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL '20, pages 3364–3374.

Ramy Baly, Georgi Karadzhov, Abdelrhman Saleh, James Glass, and Preslav Nakov. 2019. Multi-task ordinal regression for jointly predicting the trustworthiness and the leading political ideology of news media. In *Proceedings of the 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '19, pages 2109–2116, Minneapolis, MN, USA.

Alberto Barrón-Cedeno, Israa Jaradat, Giovanni Da San Martino, and Preslav Nakov. 2019. Proppy: Organizing the news based on their propagandistic content. *Information Processing & Management*, 56(5):1849–1864.

Ceren Budak, Sharad Goel, and Justin M Rao. 2016. Fair and balanced? Quantifying media bias through crowdsourced content analysis. *Public Opinion Quarterly*, 80(S1):250–271.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020a. SemEval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '20, Barcelona, Spain.

Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. 2020b. A survey on computational propaganda detection. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence and the 17th Pacific Rim International Conference on Artificial Intelligence*, IJCAI-PRICAI '20, pages 4826–4832, Yokohama, Japan.

Giovanni Da San Martino, Seunghak Yu, Alberto Barron-Cedeno, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news articles. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, EMNLP '19, pages 5636–5646, Hong Kong, China.

Kareem Darwish, Michael Aupetit, Peter Stefanov, and Preslav Nakov. 2020. Unsupervised user stance detection on Twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, ICWSM '20, pages 141–152, Atlanta, GA, USA.

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

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*, NAACL-HLT '19, pages 4171–4186, Minneapolis, MN, USA.

Yoan Dinkov, Ahmed Ali, Ivan Koychev, and Preslav Nakov. 2019. Predicting the leading political ideology of YouTube channels using acoustic, textual, and metadata information. In *Proceedings of the 20th Annual Conference of the International Speech Communication Association*, INTERSPEECH '19, pages 501–505, Graz, Austria.

Erick Elejalde, Leo Ferres, and Eelco Herder. 2018. On the nature of real and perceived bias in the mainstream media. *PloS one*, 13(3):e0193765.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030.

Doris A Graber and Johanna Dunaway. 2017. *Mass media and American politics*. SAGE Publications.

Maurício Gruppi, Benjamin D. Horne, and Sibel Adalı. 2020. NELA-GT-2019: A large multi-labelled news dataset for the study of misinformation in news articles. *arXiv preprint arXiv:2003.08444*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Benjamin D. Horne, William Dron, Sara Khedr, and Sibel Adali. 2018. Assessing the news landscape: A multi-module toolkit for evaluating the credibility

[8] http://tanbih.qcri.org/

of news. In *Proceedings of the The Web Conference*, WWW '18, pages 235–238, Lyon, France.

Benjamin D Horne, Jeppe Nørregaard, and Sibel Adalı. 2019. Different spirals of sameness: A study of content sharing in mainstream and alternative media. In *Proceedings of the International AAAI Conference on Web and Social Media*, ICWSM '19, pages 257–266, Munich, Germany.

Shanto Iyengar and Kyu S Hahn. 2009. Red media, blue media: Evidence of ideological selectivity in media use. *Journal of communication*, 59(1):19–39.

Ye Jiang, Johann Petrak, Xingyi Song, Kalina Bontcheva, and Diana Maynard. 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*, SemEval '19, pages 840–844, Minneapolis, MN, USA.

Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. 2019. SemEval-2019 Task 4: Hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, SemEval '19, pages 829–839, Minneapolis, Minnesota, USA.

Vivek Kulkarni, Junting Ye, Steven Skiena, and William Yang Wang. 2018. Multi-view models for political ideology detection of news articles. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '18, pages 3518–3527, Brussels, Belgium.

Wei-Hao Lin, Theresa Wilson, Janyce Wiebe, and Alexander Hauptmann. 2006. Which side are you on? Identifying perspectives at the document and sentence levels. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, CoNLL '06, pages 109–116.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, ACL '04, pages 271–278, Barcelona, Spain.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, EMNLP '14, pages 1532–1543, Doha, Qatar.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2018. A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, ACL '18, pages 231–240, Melbourne, Australia.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, Yejin Choi, and Paul G Allen. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, EMNLP '17, pages 2931–2937, Copenhagen, Denmark.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 '19, pages 3973–3983, Hong Kong, China.

Diego Saez-Trumper, Carlos Castillo, and Mounia Lalmas. 2013. Social media news communities: Gatekeeping, coverage, and statement bias. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, CIKM '13, page 1679–1684, San Francisco, CA, USA.

Abdelrhman Saleh, Ramy Baly, Alberto Barrón-Cedeño, Giovanni Da San Martino, Mitra Mohtarami, Preslav Nakov, and James Glass. 2019. Team QCRI-MIT at SemEval-2019 Task 4: Propaganda analysis meets hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, SemEval '19, pages 1041–1046, Minneapolis, MN, USA.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '15, pages 815–823, Boston, MA, USA.

Yanchuan Sim, Brice D. L. Acree, Justin H. Gross, and Noah A. Smith. 2013. Measuring ideological proportions in political speeches. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, EMNLP '13, pages 91–101, Seattle, Washington, USA.

Peter Stefanov, Kareem Darwish, Atanas Atanasov, and Preslav Nakov. 2020. Predicting the topical stance and political leaning of media using tweets. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL '20, pages 527–537.

Yifan Zhang, Giovanni Da San Martino, Alberto Barrón-Cedeño, Salvatore Romeo, Jisun An, Haewoon Kwak, Todor Staykovski, Israa Jaradat, Georgi Karadzhov, Ramy Baly, Kareem Darwish, James Glass, and Preslav Nakov. 2019. Tanbih: Get to know what you are reading. 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 '19, pages 223–228, Hong Kong, China.