Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 1268–1277
Marseille, 11–16 May 2020

Discovering Biased News Articles Leveraging Multiple Human Annotations

Konstantina Lazaridou,[1] **Alexander Löser,**[1] **Maria Mestre,**[2] **Felix Naumann**[3]
[1]Beuth University of Applied Sciences Berlin, Germany, [2]Factmata Ltd., London, United Kingdom, [3]Hasso Plattner Institute, University of Potsdam, Germany
konstantina.lazaridou@beuth-hochschule.de, aloeser@beuth-hochschule.de, mariarmestre@gmail.com, felix.naumann@hpi.de

Abstract

Unbiased and fair reporting is an integral part of ethical journalism. Yet, political propaganda and one-sided views can be found in the news and can cause distrust in media. Both accidental and deliberate political bias affect the readers and shape their views. We contribute to a trustworthy media ecosystem by automatically identifying politically biased news articles. We introduce novel corpora annotated by two communities, i.e., domain experts and crowd workers, and we also consider automatic article labels inferred by the newspapers' ideologies. Our goal is to compare domain experts to crowd workers and also to prove that media bias can be detected automatically. We classify news articles with a neural network and we also improve our performance in a self-supervised manner.

Keywords: bias detection, text classification, curriculum learning

1. Media Bias

Given the vast amount of news we consume on a day-to-day basis, ensuring information quality and credibility (Popat et al., 2016) becomes increasingly crucial, because we need access to accurate and reliable news stories. This way, we can form well-rounded views and make informed choices for our votes. Unfortunately, between the emergence of fake news articles, political propaganda in the media, and also hateful language around the Web, it is important to be alert and potentially show mistrust to the providers of information. We consider the following definition of media bias (based on the Oxford University Press definition): A biased news article leans towards or against a certain person or opinion by making one-sided, misleading or unfair judgements. An unbiased news article reports fair, impartial and objective information.

Media bias can be expressed in multiple ways (Saez-Trumper et al., 2013), for instance it can be present in word choices: some use the word "terrorists" vs. "freedom fighters" or "death tax"[1] vs. "inheritance tax"[2]. Even though such phenomena are present and can introduce bias in the news, reliable labelled corpora are missing to learn automatically the hidden patterns in the text. In fact, while there are relevant studies in political science (Hamborg et al., 2018), works that investigate the scope of bias (Groseclose and Milyo, 2005), how it is generated (Gentzkow and Shapiro, 2010) and others that detect it in different domains (Cohen and Ruths, 2013; Recasens et al., 2013; Iyyer et al., 2014), related work lacks automatic solutions for the binary classification task that classifies mainstream news articles as biased or unbiased.

Moreover, we have observed opinionated news pieces that are not marked as "Opinion" or "Editorial" at the beginning of the article and regardless, they use extreme political language. For instance, an article from *Right Wing News*[3] describes the Barack Obama administration as awful and another one from *Red State*[4] writes that liberals are regressive leftists with mental health issues, respectively. Even though the domain names reveal a stance in this case, other examples cannot always be captured by the commonly accepted newspaper stances (Umarova and Mustafaraj, 2019). Hence, we do not rely on predefined and commonly accepted slants of media (Patricia Aires et al., 2019), but we identify the importance of human labels for news media bias detection and introduce them in our paper.

Detecting politically toxic content on the Web can prepare and protect both news readers and online social network communities from misleading or toxic information. Journalists can also benefit from such content evaluation in order to reflect on their work. News aggregators, such as Google News, can incorporate this feature along with others (e.g., fake claim, missing citations, etc.) to facilitate the user's briefing and remove the lenses that certain news sources write their articles from. To the best of our knowledge, this is the first work that introduces news data with domain expert annotations for media bias, and compares them with crowd-sourced and silver standard (automatic) annotations as well.

Our goal is two-fold: to discover whether domain expertise is necessary for this task, and to show whether deep learning techniques can tackle such a challenging classification problem even for humans. Although our first research question might sound trivial, the complex nature of this problem and the lack of related work in computer science on news media bias leads us to investigate the differences between expert and non-expert annotations in a qualitative and quantitative

Work partially done during a full-time summer internship at Factmata.

[1] https://thenewdaily.com.au/news/national/2019/04/24/bill-shorten-death-tax/

[2] https://www.independent.co.uk/money/spend-save/hmrc-inheritance-tax-bill-rise-23-per-cent-inland-revenue-treasury\protect\discretionary{\char\hyphenchar\font}{}{}a7860626.html

[3] www.rightwingnews.com/chelsea-clinton/chelsea-clinton-attempts-burn-republicans-tweet-instead-massively-insults-michelle-obama/

[4] https://www.redstate.com/setonmotley/2018/01/03/reversing-obama-trump-protecting-thus-promoting-intellectual-property/

manner. As a second step, we focus on the automatic prediction of media bias and aim to overcome the challenge of the vague media bias definition (Hamborg et al., 2018). Our contributions include:

- We introduce novel and reliable annotated datasets for media bias detection
- We are the first to compare experts and non-expert annotators for this task
- We classify the news articles with a deep learning model and a self-supervised curriculum learning technique
- We perform an error analysis of our results for further insights of the problem

Our study is a joint work with Factmata[5], a misinformation detection company. During our collaboration, we have interacted with several native English speaking journalists that helped us assess the quality of online news, by labeling news articles, giving us feedback on labels they find helpful for media bias detection, etc. Note that it is challenging to confidently define what is biased and what is not, because bias can be perceived differently by different individuals (Groseclose and Milyo, 2005), even by experts. For instance, 80% of the journalists we collaborated with define political media bias as the act of writing the news so that they fit a specific political agenda, view or party. 15% of them believe that the bias is often inevitable and it should be explicitly declared to avoid confusion.
The rest of this paper is organized as follows: In Section 2., we examine related work. In Section 3., we introduce our datasets and Section 4. shows our data quality analysis. Section 5. describes our classification method and Section 6. presents our results. Lastly, Section 7. concludes this work and contains our ideas for future work.

2. Related Work

The problem of political bias in the news is originally and mainly tackled in **political science**, though lately it has gained attention in computer science as well. The survey by Hamborg et al. outlines the creation stages and effects of media bias (Hamborg et al., 2018). The authors also outline the different forms of selection bias that social science studies. Very few computer science works exist that study news media bias and they mainly solve related sub-problems, e.g., source, topic, sentiment and event detection. Due to the difficulty to classify articles for their bias and the lack of training data, there exist **approximations** to understand this problem, e.g., examining the outlets' quoting patterns (Niculae et al., 2015), leveraging information in social media (Zhou et al., 2011; Ribeiro et al., 2018) and the political orientation of news readers (Kulshrestha et al., 2018). Other studies reduce the complexity of the bias detection problem by focusing on the **sentence level**, namely analyzing the choices news outlets make for the statements they publish and the politicians they mention (Konstantina Lazaridou, 2016), and also the news headlines they write (Chen et al., 2018). In addition, Yano et al. annotate biased sentences in American political blogs and compare the perceived bias of the labelers to the commonly-accepted slant of the blogs (Yano et al., 2010). In contrast, we aim to classify automatically political bias in traditional news articles on the article level (noted as *spin bias* (Hamborg et al., 2018)), whose text contains mainly subtle manifestations of political viewpoints that are not encouraged as they are in political blogs.
Furthermore, reporters often change their narrative in order to focus on a certain aspect, a technique that is called *news framing*. Related work analyzes specific types of framing in the media (Morstatter et al., 2018). Another line of research performs a linguistic analysis of *hyperpartisan* (extremely biased) and fake news and shows that the latter are often politically biased (Potthast et al., 2018). **Writing style** features and readability scores are used to predict hyperpartisanship, political perspective and fake content. In general, **linguistic analyses** could reveal many interesting patterns in the text, but one might need to perform complex argument mining, opinion holder detection, or to identify direct and indirect reported speech (so that it is not attributed to the article author), etc. Political perspective detection is also studied on blogs (Lin et al., 2006; Ahmed and Xing, 2010) and news outlets (Baly et al., 2019; Patricia Aires et al., 2019). However, we focus on the binary categorization of news articles into "biased" and "unbiased", rather than on particular cases of bias, e.g., left-wing/right-wing, conservative/liberal, unreliable/trustworthy etc., Moreover, recent studies propose **textual features** for the problem of deception detection on the Web in order to find unreliable information (Volkova and Jang, 2018). The authors utilize features such as biased language lexicons, connotation frames, writing style, etc. Opposed to this setting, we do not perform any cumbersome feature engineering, but we rely only on the content of the articles we classify. To the best of our knowledge, there is not an existing automatic solution for classifying a news article in a binary manner as biased or unbiased, mainly due to the unavailability of reliable document-level labels by trustworthy annotators. Another reason is the noise of the existing labels inferred from the commonly accepted stance of the newspapers (Umarova and Mustafaraj, 2019). These inferred assumptions could potentially change over time due to trends or new owners and reporters joining the news outlets. In contrast, human labels are more reliable and potentially explainable, e.g., by looking into the annotator agreement or the notes annotators leave while labeling. In this work, we focus only on mainstream news media without utilizing textual features and predefined media slants, but we guide and improve our classification model by applying curriculum learning (Bengio et al., 2009). This technique has been shown to improve the classification performance and the training process in machine learning. It is also reported to outperform non-curriculum approaches in multiple tasks, such as language modeling, especially when the task is particularly challenging like ours (Weinshall et al., 2018).

3. Novel Datasets

In this section we describe how we collect our political news datasets on arbitrary topics in 2015–2018.

[5] https://factmata.com

				Classes		
Dataset	Articles	Annotations/Article	Labels	Biased	Unbiased	Newspapers
Experts (E)	1,154	3	0, 1	523	631	306
Non-experts (NE)	2,993	3	1, 2, 3, 4, 5	1197	1230	961
Publishers (P)	750,000	1	0, 1	375,000	375,000	1194

Table 1: Data characteristics: Number of news articles in each collection, number of annotations per article, labels, number of articles in each class and number of unique newspapers in each dataset.

3.1. News Corpora for Bias Detection

All datasets are presented in Table 1. We gathered political news articles from a broad variety of news sources in terms of size and credibility for our annotation task.[6] Additionally to these humanly labeled articles (*E* and *NE*), we use the training data given to the participants of the Semeval 2019 task for hyperpartisanship detection (denoted as *P*) to compare our performance against it (Kiesel et al., 2019). These publisher-based labels are produced based on newspaper credibility scores.

Articles annotated by journalists. As shown in Table 1, this is a rather small collection (*E*). However, due to the experience of the annotators in their field and their ability to identify one-sided text even in cases where bias is very subtle, we hypothesize that this dataset is very valuable. This set of news articles is included in the non-expert data as well (*NE*), in order to facilitate their comparison. The platform that was used is an internal annotation tool of Factmata, where the users (eight journalists) were asked to read a set of political news articles and mark at least one biased or unbiased text snippet that they find in each article, following the bias definition in Section 1. (i.e., the author is favoring or discriminating a certain view or person). The labelers were asked to identify the bias of the overall article and then highlight the evidence for their decision. By extension, the annotations can be words, sentences, paragraphs or entire documents. We chose this setting, because these low-level annotations can give more concrete evidence of bias and can be used as ground truth for explaining our model in the future (Arras et al., 2017).

We propagate these fine-grained labels to the article level and we assume that each article that contains at least one annotated biased (or unbiased) sentence is biased (or unbiased respectively). We exclude articles that contain both biased and unbiased marked text. This filter prunes less than 1% of the data, because the journalists were asked to not annotate each document exhaustively. It is obvious that regardless the annotations, a biased article can contain neutral text as well (and vice versa). We manually examined the excluded articles and we observed that sometimes in these cases the text contains the relevant facts, but also a few opinionated words that one might identify as biased. It also occurs that such articles are biased towards a given perspective, but they are well-written and cite the appropriate sources. We regard them as unclear, but we are interested in gaining insights into these potentially controversial news pieces in our future work.

Articles annotated by the crowd. The next dataset consists of annotations from our two crowd-sourcing tasks for media bias detection, launched in the Amazon Mechanical Turk[7] (1,979 documents) and the Figure Eight[8] (1,014 articles) platforms. Note that the Figure Eight dataset was originally introduced in 2018 (Vincent and Mestre, 2018), though in this work we consider the full dataset, instead of the proposed filtered version based on an in-house evaluation of the data. In both datasets, the crowd workers evaluated each article using a score range similar to related work (Yano et al., 2010), where 1 meant “unbiased" and 5 signified “biased". Similarly with the experts, they were asked to follow the media bias definition in Section 1. and read the full article before they annotate. Both the crowd and the experts were asked to be mindful of bias manifestations, such as loaded or subjective language, opinionated text, one-sided claims, or unsupported arguments. As we can observe in Table 1, the combination of these two non-expert (*NE*) data collections contains almost 3,000 news articles labeled for their political bias. We have combined the annotations from these two tasks into one unified dataset. The expert and non-expert document collections are available via our industry collaborator for further details and research purposes.

3.2. Data Preprocessing

In this section, we explain how we aggregate and transform our datasets.

Article transformation. There are at least three annotations per article in *E* and *NE*, and the class distribution in each case is fairly balanced. In order to aggregate the labels of multiple annotators for each article, we apply the *Dawid Skene* algorithm (Dawid and Skene, 1979), specifically an optimized variation of it (Sinha et al., 2018). This model produces one final label for each document and it improves on simpler methods, because it considers the annotators' bias and competence. It is assumed that each worker corresponds to a confusion matrix that shows the joint probability distribution over correct and reported labels. The correct labels are initialized with the *Majority Vote* method, which outputs the label that was reported most often. For a N-way classification task (in our case $N = 2$), a worker w and a

[6] Example news sources: AbcBusinessNews, Associated Press, Albuquerque Journal, Baptist News Global, BBC, Breitbart, Chicago Reporter, Circa News, CNN, CounterCurrents, Daily Banter, Ethics and Public Policy Center, Fair, Federalist Press, Fox Business, Free Beacon, Greensboro, Guardian, Heavy, InfoWars, Intrepid Report, In These Times, Lima Charlie News, MotherJones, MSNBC, NBC News, NewsMax, New York Times, Occupy, OpsLens, Political Insider, Poynter Institute, Raw Story, Real News Network, Reuters, San Jose Mercury News, Seattle Times, Slate, Times of India, Townhall, Upworthy, Valley News, Vox, Washington Blade, 21st Century Wire, The Whim.

[7] https://www.mturk.com/

[8] https://www.figure-eight.com/

data instance d, the Dawid-Skene assumption is as follows:

$$P(X_{wd} = l) = p^*_{wlx_d}$$

where X_{wd} is the random variable that models the reported label l of annotator w and for the document d, and all X_{wd} are mutually independent. After generating one annotation per article, we still face the challenge that *E* contains binary labels, but *NE* corresponds to a multi-class classification setting. For this purpose, we binarize the non-expert data, following the literature in similar tasks where five star ranges were used (Maas et al., 2011). We take into account only the two ends of the scale, namely only the highly polarized text. That is, we consider the articles with bias score 1 and 2 as unbiased (negative class), and the ones with bias score 4 and 5 as biased (positive class). Similarly to *E*, we exclude ambiguously labeled data (bias score is 3).

Unambiguous test set for media bias detection. We construct a reliable and independent of our training data test set in order to compare the achieved classification performance with training data labeled by different communities. We use a subset of the common articles that are annotated by experts and non-experts, namely all articles in *E*. We further consider the subset of articles that are marked with the same label both by the experts and the crowd, because we hypothesize that these articles have low uncertainty and controversy regarding the underlying media bias. From this unambiguous dataset, we randomly sample 40% of it and use it as our final test set. We leave the rest 60% in *E* and *NE* respectively. We do so in order to maintain our training data sufficiently large, given that in our experiments we remove from the training sets any article that appears also in the test set. Hence with this setting, our training data contain "diverse" articles, whose labels might or might not be the same in *E* and *NE*.

4. Label Quality Assessment

In this section, we describe our annotation analysis as an effort to determine the quality of the datasets and improve our classification results later on.

4.1. Per-dataset Agreement

As a first step to examine the quality of the human labels, we measure the inter-annotator agreement (*ITA*) within each collection. That is, we calculate the agreement for the expert dataset, the Figure Eight dataset and the MTurk dataset separately. Note that for this experiment we consider the original labels in the raw data, without binarizing them first (we transform the labels as described in Section 3.2. only later on for machine learning purposes). We chose Krippendorff's α coefficient[9], which is independent of the sample size, the categories, and numbers of annotators and measurement levels. Krippendorff's α for a text document is defined as follows:

$$\alpha = \frac{p_a - p_e}{1 - p_e}$$

[9] https://en.wikipedia.org/wiki/Krippendorff%27s_alpha

Dataset	*ITA*
Crowd workers (Figure Eight)	0.21
Experts (Journalists)	0.59
Crowd workers (MTurk)	0.66

Table 2: Inter-annotator agreement (Krippendorff's α) for each of the three humanly labeled datasets.

where p_a is the weighted percent agreement and p_e to the weighted percent chance agreement. According to this metric, the documents and the agreement scores assigned to them are statistically unrelated. When $\alpha = 1$, this indicates perfect reliability and when $\alpha = 0$, there is absence of reliability. Moreover, α is zero when disagreements are systematic and exceed what can be expected by chance.

We present our findings in Table 2. Considering how challenging the given problem is, we observe the expert (*E*) and MTurk annotators to agree sufficiently well internally in each collection. However, the data produced via Figure Eight seem more ambiguous. Chronologically, we have performed these annotations tasks starting with Figure Eight, continuing with the journalists and then completing our study with MTurk. That is why the differences in the agreement could be justified due to the continuous improvement of our instructions to the annotators, which potentially makes the annotations' quality higher at the later rounds in contrast to the earlier ones. For instance, we discovered that we had to explicitly emphasize to all annotators the difference between when a reporter's words and viewpoints are toxic themselves, to when a politically toxic event or statement is reported, and that we are only interested in the first case.

Furthermore, the labeled dataset from Figure Eight was introduced earlier (Vincent and Mestre, 2018), where an in-house gold standard dataset based on fact-checking was used to evaluate the workers and disqualify unreliable ones. In our study we consider the full dataset (thus, we see a lower inter-annotator agreement), in order to maintain a more generalized setting without constraints. Note that both crowd-sourced datasets use a numerical range for the bias score. We leverage the numerical distance between the labels when computing the *ITA*, which is not possible in a binary setting, e.g., in the expert dataset. Taking this range into account, we have significantly improved the inter-annotator agreement (from 0.14 to 0.21 in Figure Eight and from 0.44 to 0.66 in MTurk), where both original agreement scores are lower than in *E*.

4.2. Cross-dataset Agreement

To investigate whether media expertise is necessary for our task and ultimately which annotator group is more appropriate to solve our problem, we compute the annotator agreement between *E* and *NE*. For the crowd-sourced data, the transformed annotations to binary labels are used as described in Section 3.2.. We apply a well-established method for expert versus non-expert analysis in natural language processing tasks (Snow et al., 2008), using the articles that both *E* and *NE* annotated. The authors calculate how (non-) experts perform within their community and against all involved annotators (experts and non-experts combined). Given two communities A and B, for every individual a_i

in A, they compute the *ITA* with all the individuals b_j in B and then average the results. In the following step, they average across all individuals a_i in order to obtain how well A agrees with B in total. The authors use the Pearson correlation coefficient (*PCC*) as agreement metric. Given two vectors (the labels of two different annotators), the computed *PCC* has a value between 1 (positive correlation) and -1 (negative correlation). For our task, this agreement metric is not appropriate, because not all user pairs annotated exactly the same amount of articles and this makes *PCC* not work as expected: it yields a high score when two annotators have many common articles, and very low score (close to zero) when the shared articles are few. We have worked with a limited number of eight journalists, as it is cumbersome and expensive to obtain domain expert annotations, but the crowd-sourcing platforms are generally low-cost and employ a very high number of annotators for their tasks (in our case eighty). Thus, the non-experts have annotated generally more articles and also more articles in common with each other – the latter could make their agreement scores more robust. We apply a simpler method instead of *PCC*, i.e., the percentage of times that two annotators agreed on the article bias.

Our findings are presented in Table 3. Surprisingly, the expert community and the crowd-workers appear to agree on what is biased and what is not at approximately 70% of the time. Thus, in the majority of the articles the individuals in *E* and *NE* recognize the evidence in the text to mark it as biased or unbiased. We hypothesize that in the majority of the agreement cases the articles are either very obviously hyperpartisan or very fair and balanced news, and potentially the disagreement occurs when the article topics are more controversial and ambiguous. Interestingly, the STDEV in *E* vs. *All* is much lower than in *NE* vs. *All*, which can be an indicator of the consistency and reliability of the journalists. Thus, given the lower variance, one might not need as many expert annotators as crowd workers to obtain a high quality media bias detection dataset. Moreover, journalists do not agree with each other significantly more than non-experts agree with one another. This could be potentially explained by the fact that media bias can be a very sensitive and often times subjective topic for journalists. Therefore, so far we observe unexpected yet not entirely conclusive results regarding the superiority of either annotator group.

Compared sets	Agreement %	STDEV %
E vs. *NE*	73.68	17.13
E vs. *E*	65.46	15.46
E vs. *All*	67.39	7.87
NE vs. *NE*	64.76	19.51
NE vs. *All*	64.32	20.3

Table 3: Inter-annotator agreement and standard deviation based on the method of Snow et al. (Snow et al., 2008). *E* refers to the experts, *NE* to the non-experts and and *All* to both. For the crowd-sourced data, the binarized annotations are used as described in Section 3.2.

5. Article Classification

In this section, we describe our approach to detect media bias automatically and how we apply a curriculum learning technique to improve our results.

5.1. Baseline Method

We use the FastText classifier (Joulin et al., 2016), a basic neural network that uses averaged n-gram features, well-known for competitive results to state-of-the-art approaches for text classification. We run all our experiments with the learning rate set to 0.1 and for 500 epochs.

5.2. Curriculum Learning Method

To enhance our baseline model, we leverage the data quality assessment we performed in the Section 4., and apply a *curriculum learning* approach, which is based on *transfer learning*: Given a target classification task T_1 and an external one T_2, transfer learning techniques that solve T_1 could leverage information derived by T_2 in different ways. For instance, one can use word embeddings or losses of output layers trained on T_2, or take an entire network designed for T_2 and train it on T_1 to improve the classification performance. Unlike traditional transfer learning approaches, our external information is not provided by another classifier, dataset or task. In contrast, it is derived by the humans that share our mission to fight misinformation on the Web and contribute to our task by labeling our political news articles. Ultimately, using their wisdom, we aim to guide our classifier during training with some initial data instances ("easy to learn examples") and perform better in the next steps.

We follow the definition of *curriculum* as introduced by Bengio et al. (Bengio et al., 2009), i.e., sorting the training examples from "easy" to "difficult" and introducing them to our classifier in this order during training to avoid confusing the learner. This method can not only speed up the training process, but it can improve the classification results and model generalization as well. The authors perform experiments on shape recognition and language modeling in their work. For the latter task (which is more relevant to ours), the curriculum learning strategy was to grow the vocabulary size gradually, i.e., starting from the most popular words in a Wikipedia corpus and then considering more words in each training pass.

Our learning difficulty definition. In our proposed approach, we leverage our previous agreement analysis and build a curriculum that stems from the quality of the article annotations. That is, we compute the inter-annotator agreement (*ITA*) in *E* and *NE* with the Krippendorff's alpha coefficient as shown in Section 4.1. We consider the agreement score to be the learning difficulty of an article. This choice is based on the assumption that an "easy" article is an article that causes very low to no disagreement between its annotators regarding its bias. We hypothesize that these news pieces are either very objective or very subjective, and hence this makes the decisions of the annotators simpler. On the other side, newspaper articles with high label disagreement may indicate controversy and potentially contain a mixture of facts and opinionated words. We leave these difficult-to-learn examples to be given to our model after the clearer examples have been introduced. We split the training data into 10 parts, namely we first consider the top-10% of the documents with the highest agreement score, then the top-20% and so on and so forth. We build a classifier with

each of these data chunks and every time we load the latest calculated weights from the previously trained model. We then fit the current training set and predict the media bias of our test set.

Our technique is also similar to the *stochastic curriculum learning* definition (Weinshall et al., 2018), which is a variation of stochastic gradient descent, where the model imports training data instances gradually based on their difficulty score. In this case, the authors define their curriculum without the presence of human knowledge. They transfer information from another learner, namely, they consider the difficulty of each of their data points to be the confidence (margin) of a support vector machine classifier (*SVM*) trained for the same task. Hence, the authors rank the documents based on the results of another model and feed them progressively into their own model in descending order of their difficulty. They also try different scheduling mechanisms to sample the training data, namely *fixed* (similar to our approach) and *adaptive* (based on the loss of each step).

Evaluation setting. Since the annotator agreement results for the Figure Eight dataset were not satisfactory, we use only the MTurk dataset for the rest of our experiments. Thus, when we refer to the non-expert annotations, only the dataset from MTurk is considered. We have actually trained our model on the Figure Eight dataset and the performance was similar to a random decision – we do not report the detailed results here. Hence, we concluded that it is necessary for the inter-annotator agreement to be at least 60% for our task and the Figure Eight labels did not achieve it. Related work also reports similar results (0.55 Cohen Kappa score) for crowd-sourcing biased sentences in blogs. Furthermore, even though the precision achieved with the Figure Eight dataset in the work of Vincent (Vincent and Mestre, 2018) is promising (approx. 70% on their own test set), the in-house manual improvements during training and testing that the authors perform raise the question of generalization potential of their approach. Thus, we leave investigating this data for future research.We show the size of our training and test sets in Table 4. As mentioned earlier in Section 3.2., our test set is balanced and it consists of a random sample of the news articles for which both the domain experts and the crowd workers agreed on their labels in order to eliminate noise. After removing these 237 articles from the training data, there are 759 articles in *E* and 1,805 in *NE* remaining for our model to learn (*P* does not exhibit an overlap with *E* and *NE*). Note that a similar setting was used at the Semeval competition, where the unknown test set was also a small balanced (and crowd-sourced) dataset of 645 news articles. We also prefer this predefined unseen test set instead of cross validation (which is appropriate for small datasets like ours), because we can maintain the same test set across all our experiments with different training datasets. Especially for the Semeval data that we compare against, cross-validation would not work, as we only aim to test on humanly annotated documents.

6. Results

In this section we describe our experimental evaluation and we show the qualitative results of our error analysis.

Training sets			Test set
E	*NE*	*P*	$\subset(E \cap NE)$
759	1,805	750,000	237

Table 4: Article sizes of our three different training sets and our unambiguous test data.

Training	Precision	Recall	F-1
E	0.90	0.89	0.89
NE	0.85	0.89	0.87
E_c	0.93	0.95	0.93
NE_c	0.79	0.86	0.82

Table 5: Classification results of our model trained with: expert data, non-expert data, expert data with curriculum learning and non-expert data with curriculum learning. Our test set is a sample of the articles where both experts and non-experts agree.

6.1. Domain Expertise Stands out

We use the manually expert and non-expert annotated articles (*E* and *NE* respectively) for training a FastText classifier. In the first two lines of Table 5 we can see that the articles annotated by journalists are a more appropriate dataset for this task, because when our model is trained with it, it achieves significantly higher precision. The model trained with crowd-sourced labeled articles constitutes a promising dataset that achieves competitive results with the expert model, though it does not outperform the performance of the model trained with *E*. Note that even though the consensus in MTurk is higher than in the expert data (see Table 2), the prediction power of MTurk is lower. Thus, higher inter-annotator agreement does not automatically lead to higher classification results in this case. In the following lines we see the classification results of the same models, but this time trained incrementally with a curriculum created based on the learning difficulty of each data instance. The achieved F-measure with *E_c* is significantly higher than the one with *E*, namely 93%. Note that recent related work on similar tasks (Potthast et al., 2018) that uses linguistic features of news articles achieves a maximum of 86% precision for hyperpartisanship and 75% precision for political orientation classification.

Unfortunately, the curriculum constructed by the knowledge of the crowd is not as useful for this task as the one by experts. In fact, it worsens the performance of *NE* by decreasing the achieved precision from 85% to 79%. Note that the training dataset constructed by crowd workers is more than twice the size of the one by journalists and overall it still shows a lower F-1 measure for our task, with or without curriculum learning. We hypothesize that this outcome signifies the limits of mass labeling in crowd-sourcing platforms for tasks that are often not clearly defined and easily solvable even by humans, e.g., bias, irony and sarcasm detection. Furthermore, we also performed experiments with a dense feed forward neural network and a network with long-short memory units (*LSTM*) that we do not report in detail here. Our results were not as satisfactory as with FastText (approximately 20% worse). We hypothesize that these networks are potentially too big and too complex for our small humanly labeled datasets (which is why transformer

Training	Precision	Recall	F-1
E_c	0.93	0.95	0.93
E_rc	0.85	0.90	0.88
NE_c	0.79	0.86	0.82
NE_rc	0.78	0.87	0.83
P	0.54	0.89	0.67

Table 6: Comparison of our models (*E_c* and *NE_c*) with anti-curriculum learning (*E_rc* and *NE_rc*) and with learning from automatically labelled data (*P*). Our test set is a sample of the articles where both experts and non-experts agree.

architectures would also likely not work). Traditional news articles are also less noisy datasets in contrast to text that is user generated, and thus a word-based input is appropriate for our task. For future work, we are interested in applying attention mechanisms that might capture specific biased terms in the text.

6.2. Bias Detection Requires Expert Curriculum

We compare our proposed solution to different methods in Table 6. In order to confirm the usefulness of an expert curriculum, we compare our approach with an *anti-curriculum* approach. Namely, we rank our training data instances in an ascending order of their learning difficulty as defined in Section 5.2.. In this way, we introduce the most ambiguous and hard to learn examples to our classifier first, and then proceed with the rest of the training data, completing the learning process with the easiest examples. We show in Table 6 that, as expected, this "reverse" curriculum technique (*E_rc*) worsens the results of our expert-based model significantly. In addition, it produces almost the same outcome for the non-expert data (*NE_rc*). We hypothesize that for this reason the labels of the crowd are not of the same potential as the ones by journalists. That is, the non-expert consensus for a given article does not provide additional intuition or help to a media bias detector.

In Figure 1 we show how our precision increases while we increase the training set size using *E_c*, *E_rc*, *NE_c* and *NE_rc*. We observe that *E_c* outperforms the rest during the whole training process, and it starts approximately at the same precision value as *NE_c* does. It is remarkable that only the top 20% of the expert data with the lowest learning difficulty can already achieve 80% precision. Furthermore, all four models improve as the training set size increases, however the curves of *NE_c* and *NE_rc* almost overlap. This indicates that a crowd curriculum does not prevail over its anti-curriculum version, and thus it is not as helpful to the learner as the expert-based one. Lastly, we see a significant difference between *E_c* and *E_rc* both in the starting point and during training.

6.3. Quality is More Important than Quantity

We perform an additional comparison of our approach to a model trained with automatically labelled articles for their media bias. As briefly mentioned in Section 3., we consider articles with inferred publisher-based bias from a Semeval competition (Kiesel et al., 2019). In this dataset (*P*), each document is marked as hyperpartisan (or not) if the news

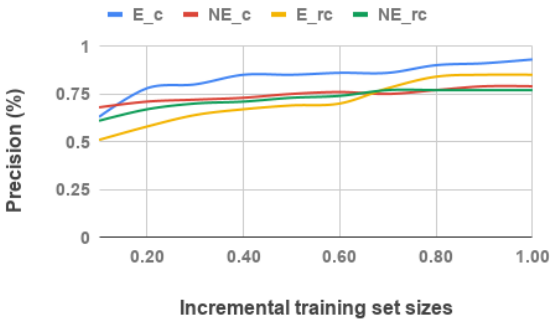

Figure 1: Achieved precision by our model trained with expert curriculum (*E_c*), non-expert curriculum (*NE_c*), compared to the respective reverse settings, *E_rc* and *NE_rc*. We train incrementally with step=10%, starting with the top 10% of the training set that is the "easiest" (*c*) or "hardest" (*rc*) to learn.

outlet where the article originates from is considered extremely politically biased (or not). As shown in Table 1, this dataset is considerably larger than ours. Note that a few hundreds of annotated articles by the crowd were included in the Semeval training set by the organizers, which we removed from *P*, because we aim to compare our small manually labeled training datasets to a massive silver standard dataset.

The comparative results are shown at the bottom in Table 6. It is evident that small amounts of human labels and domain expertise are more essential for our task than the size of training data, which is not true for every machine learning problem. Large amounts of weak labels are sufficient for other tasks, such as sentiment classification (Deriu et al., 2017). Other similar works that miss correct labels include tracking and matching individuals in images with transfer learning (Peng et al., 2016). The outcome can also be justified, because we consider the Semeval dataset to address a slightly different task, namely the prediction of the newspaper's bias and not the article's (similarly to Aires et al., where domain level labels are used (Patricia Aires et al., 2019)). It is expected that the Semeval dataset can achieve competitive recall values (almost 90%) due to its very large size, but the precision is still suffering from the uncertain quality in the training set. Our manual qualitative examination shows that there is significant noise in Semeval the data, which is on par with the results (60–70% classification accuracy) of recent studies based on this dataset (Saleh et al., 2019). For instance, every news article by *Pjmedia* is considered hyperpartisan in this dataset, but not all articles from this news source in the MTurk dataset are labeled as such. Note that although our test set is small, after comparing our approach with the classifier trained on *P*, we consider our results indeed significant. That is, we assume that if our test set was easy to classify, then the baseline with the Semeval data would be able to outperform or at least compete with our proposed expert-based approach.

6.4. Qualitative Analyses Bring Further Insights

In this section we analyze the errors of our expert curriculum model (*E_c*) and also apply it to a new dataset.

False predictions. Approximately 9% of our predictions are incorrect. Over 50% of the articles that are misclassified contain **loaded language**. Heavy words can be either the journalist's (an article calls Donald Trump "misogynist") or could describe a sensitive topic (the same article is discussing "sexual assault"). Hence, sentiment detection alone would be a rather inconclusive approach, because such words are not always chosen by the journalist, but are often contained in cited text (this is one of the multiple reasons that sentiment analysis performs poorly on the news corpora (Hamborg et al., 2018)). Moreover, in 60% of the errors the articles contain both facts and opinions. Some of them are essentially opinion pieces with factual information and verified sources, but are **disguised columns**, i.e., there is no declaration of this in any part of the news article page. Even though they are annotated correctly and they are almost all classified correctly by the model, there is still a very small set that is very hard to classify automatically. In addition, in around 30% of the errors we observe humor and satire in the text, thus a filter or another model could be used to avoid such cases.

A very interesting error class with approximately 60% of errors appears when essentially the **news topic is a politician**, and not a political event. Among these errors, over 85% of them are false positives and the rest are false negatives. Such articles generally report a politician's statement or action, and at the same time describe them with endorsement or criticism. An example is the article of *MSNBC*, where the journalist describes Michelle Obama's standpoints on Donald Trump's taped comments about women. This article is a representative error, because it has somewhat subjective tone ("Michelle Obama *slammed* Donald Trump"), discusses a sensitive subject and it is about two politicians.

A very challenging task for our model is to distinguish the presence of bias when the article is about very **sensitive topics**, e.g., incidents of racism, sexual assault, terrorism, brutal crimes. These errors (13%) are all false positive predictions, which indicates that loaded language can sometimes lead the model to confuse tragic news stories with biased reporting. Note that this is not a rule, because we also classify relevant unbiased articles correctly (e.g., an article in *Circa News* about the domestic terrorism attack in Charlottesville was a true negative). Furthermore, in about 10% of all errors, the article author is using first-person pronouns, which could be discovered with claim/argument mining. First-person expressions could serve as an indicator that an article contains the author's/newspaper's subjective point of view (Généreux and Santini, 2007) or that it is an editorial (Bonyadi, 2010).

Results on independent dataset. We additionally apply our model to a small recent set of news articles from the New York Times. We use the newspaper's *Most Popular API* to get the most read articles in mid August 2019. Out of the 17 articles in this test set, our model classified 13 as unbiased (including an opinion article that we missed), and only four of them were classified as biased. Among these four, one article is an opinion piece and another one is self-help guide giving relevant professional opinions, which justifies the decision of our model. The other two articles are about brutal crimes in Afghanistan and New York, respectively, and we consider them falsely classified as biased. The first article about a suicide bomber that killed dozens of people in the capital city of Afghanistan describes the tragic event with factual reporting. However, the language is somewhat loaded (mainly due to the nature of the news story) and the title is described by one commenter as too dramatic. Similarly to our error analysis of our own test set, we see that such tragic event reports are harder to classify correctly. Moreover, in the article about a crime committed by a police officer in the New York region, we observe only factual and fair reporting. Thus we regard this as false positive prediction as well. According to MediaBiasFactCheck[10], the New York Times is a highly factual and reliable unbiased source, that occasionally publishes articles with loaded language that moderately favors liberal views. We find our qualitative study to be on par with MediaBiasFactCheck, because our model labels the majority of the articles unbiased, captures almost all the opinion pieces (which are explicitly declared and do not belong to our focus) and understandably misclassifies the articles on hard-to-classify topics due to the presence of emotional words.

7. Summary and Future Work

In this paper, we tackle the problem of media bias detection in the news. We introduce two novel humanly labeled article sets and use them to build very competitive deep learning models for our task. Our work is the first to consider and compare human labels (by domain experts and crowd-source workers) to automatically derived labels for media bias detection. We classify news articles successfully for their bias and also give human knowledge to our model as a *curriculum*, by introducing the articles incrementally during training. Our conclusion is that human labels are more suitable than automatic labels for this task, with both models trained on crowd and expert data respectively achieving higher F-1 scores. The expert knowledge can be used in the form of a curriculum to boost the classification performance further, e.g., a model trained with the top-20% articles with the highest consensus among experts can already achieve 80% precision. Moreover, we show that the inter-annotator agreement score for our task should be at least 60% and that the amount of training data is not as influential as its quality. We also contribute further insights with our manual error interpretation and discover challenging corner cases to be aware when annotating or classifying new media bias. In the future, we aim to shed more light into our ambiguous human annotations (articles marked with score=3/5 in the non-expert data and articles with biased and unbiased snippets in the expert data). This set of articles could become a very difficult and interesting test set for our task, or a training set for controversy detection in the news. Moreover, we plan to experiment with stricter learning difficulty scores, e.g., the global annotator agreement in all collections instead of the internal agreement within each collection. We intuit that this could result to an even more robust and powerful curriculum.

[10] https://mediabiasfactcheck.com/new-york-times/

8. Bibliographical References

Ahmed, A. and Xing, E. P. (2010). Staying informed: supervised and semi-supervised multi-view topical analysis of ideological perspective. In Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics.

Arras, L., Montavon, G., Müller, K.-R., and Samek, W. (2017). Explaining recurrent neural network predictions in sentiment analysis. In Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis.

Baly, R., Karadzhov, G., Saleh, A., Glass, J., and Nakov, P. (2019). Multi-task ordinal regression for jointly predicting the trustworthiness and the leading political ideology of news media. *Computing Research Repository*.

Bengio, Y., Louradour, J., Collobert, R., and Weston, J. (2009). Curriculum learning. In Proceedings of the 26th ACM Annual International Conference on Machine Learning, pages 41–48.

Bonyadi, A. (2010). The rhetorical properties of the schematic structures of newspaper editorials: A comparative study of english and persian editorials. *Discourse & Communication*, 4(4):323–342.

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, pages 79–88.

Cohen, R. and Ruths, D. (2013). Classifying political orientation on twitter: It's not easy! In Proceedings of the International AAAI Conference on Weblogs and Social Media. Association for the Advancement of Artificial Intelligence.

Dawid, A. P. and Skene, A. M. (1979). Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28.

Deriu, J., Lucchi, A., De Luca, V., Severyn, A., Müller, S., Cieliebak, M., Hofmann, T., and Jaggi, M. (2017). Leveraging large amounts of weakly supervised data for multi-language sentiment classification. In Proceedings of the International Conference on World Wide Web.

Généreux, M. and Santini, M. (2007). Exploring the use of linguistic features in sentiment analysis. In Proceedings of the 4th International Corpus Linguistics Conference.

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

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

Hamborg, F., Donnay, K., and Gipp, B. (2018). Automated identification of media bias in news articles: an interdisciplinary literature review. *International Journal on Digital Libraries*.

Iyyer, M., Enns, P., Boyd-Graber, J., and Resnik, P. (2014). Political ideology detection using recursive neural networks. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pages 1113–1122.

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2016). Bag of tricks for efficient text classification. *Computing Research Repository*.

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, pages 829–839.

Konstantina Lazaridou, R. K. (2016). Identifying political bias in news articles. *Special Issue of the Bulletin of the IEEE Technical Committee on Digital Libraries, Doctoral Consortium of the 20th International Conference on Theory and Practice of Digital Libraries*, 12(3).

Kulshrestha, J., Eslami, M., Messias, J., Zafar, M. B., Ghosh, S., Gummadi, K. P., and Karahalios, K. (2018). Search bias quantification: investigating political bias in social media and web search. *Information Retrieval Journal*, 22(1):1–40.

Lin, W.-H., Wilson, T., Wiebe, J., and Hauptmann, A. (2006). Which side are you on?: Identifying perspectives at the document and sentence levels. In Proceedings of the 10th Conference on Computational Natural Language Learning, pages 109–116.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. (2011). Learning word vectors for sentiment analysis. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 142–150.

Morstatter, F., Wu, L., Yavanoglu, U., Corman, S. R., and Liu, H. (2018). Identifying framing bias in online news. *ACM Transactions on Social Computing*, 1(2):5.

Niculae, V., Suen, C., Zhang, J., Danescu-Niculescu-Mizil, C., and Leskovec, J. (2015). Quotus: The structure of political media coverage as revealed by quoting patterns. In Proceedings of the World Wide Web Conference. Association for Computing Machinery.

Patricia Aires, V., G. Nakamura, F., and F. Nakamura, E. (2019). A link-based approach to detect media bias in news websites. In Proceedings of the 30th World Wide Web Conference, Companion, pages 742–745. Association for Computing Machinery.

Peng, P., Xiang, T., Wang, Y., Pontil, M., Gong, S., Huang, T., and Tian, Y. (2016). Unsupervised cross-dataset transfer learning for person re-identification. In Proceedings of the IEEE conference on computer vision and pattern recognition.

Popat, K., Mukherjee, S., Strötgen, J., and Weikum, G. (2016). Credibility assessment of textual claims on the web. In Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, pages 2173–2178.

Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., and Stein, B. (2018). A stylometric inquiry into hyperpartisan and fake news. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, pages 231–240.

Recasens, M., Danescu-Niculescu-Mizil, C., and Jurafsky, D. (2013). Linguistic models for analyzing and detecting biased language. In Proceedings of the 51st Annual

Meeting of the Association for Computational Linguistics, pages 1650–1659.

Ribeiro, F. N., Henrique, L., Benevenuto, F., Chakraborty, A., Kulshrestha, J., Babaei, M., and Gummadi, K. P. (2018). Media bias monitor: Quantifying biases of social media news outlets at large-scale. In Proceedings of the 12th International AAAI Conference on Web and Social Media, pages 290–299.

Saez-Trumper, D., Castillo, C., and Lalmas, M. (2013). Social media news communities: gatekeeping, coverage, and statement bias. In Proceedings of the 22nd ACM International Conference on Information and Knowledge management, pages 1679–1684.

Saleh, A., Baly, R., Barrón-Cedeño, A., Da San Martino, G., Mohtarami, M., Nakov, P., and Glass, J. (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, pages 1041–1046.

Sinha, V. B., Rao, S., and Balasubramanian, V. N. (2018). Fast dawid-skene: A fast vote aggregation scheme for sentiment classification. In Proceedings of the 7th Workshop on Issues of Sentiment Discovery and Opinion Mining.

Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. Y. (2008). Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In Proceedings of the 13th Conference on Empirical Methods in Natural Language Processing, pages 254–263.

Umarova, K. and Mustafaraj, E. (2019). How partisanship and perceived political bias affect wikipedia entries of news sources. In Proceedings of the 30th World Wide Web Conference, Companion, pages 1248–1253.

Vincent, E. and Mestre, M. (2018). Crowdsourced measure of news articles bias: Assessing contributors' reliability. In Proceedings of the 1st Workshop on Subjectivity, Ambiguity and Disagreement in Crowdsourcing, and Short Paper Proceedings of the 1st Workshop on Disentangling the Relation Between Crowdsourcing and Bias Management, pages 1–10.

Volkova, S. and Jang, J. Y. (2018). Misleading or falsification: Inferring deceptive strategies and types in online news and social media. In Proceedings of the 27th World Wide Web Conference, Companion, pages 575–583.

Weinshall, D., Cohen, G., and Amir, D. (2018). Curriculum learning by transfer learning: Theory and experiments with deep networks. In Proceedings of the 35th International Conference on Machine Learning, pages 5235–5243.

Yano, T., Resnik, P., and Smith, N. A. (2010). Shedding (a thousand points of) light on biased language. In Proceedings of the NAACL HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk. Association for Computational Linguistics.

Zhou, D. X., Resnick, P., and Mei, Q. (2011). Classifying the political leaning of news articles and users from user votes. In Proceedings of the International AAAI Conference on Weblogs and Social Media. Association for the Advancement of Artificial Intelligence.