Improving Linguistic Bias Detection in Wikipedia using Cross-Domain Adaptive Pre-Training

Karthic Madanagopal
karthic11@tamu.edu
Texas A&M University
College Station, Texas, USA

James Caverlee
caverlee@tamu.edu
Texas A&M University
College Station, Texas, USA

ABSTRACT

Wikipedia is a collective intelligence platform that helps contributors to collaborate efficiently for creating and disseminating knowledge and content. A key guiding principle of Wikipedia is to maintain a neutral point of view (NPOV), which can be challenging for new contributors and experienced editors alike. Hence, several previous studies have proposed automated systems to detect biased statements on Wikipedia with mixed results. In this paper, we investigate the potential of cross-domain pre-training to learn bias features from multiple sources, including Wikipedia, news articles, and ideological statements from political figures in an effort to learn richer cross-domain indicators of bias that may be missed by existing methods. Concretely, we study the effectiveness of bias detection via cross-domain pre-training of deep transformer models. We find that the cross-domain bias classifier with continually pre-trained RoBERTa model achieves a precision of 89% with an F1 score of 87%, and can detect subtle forms of bias with higher accuracy than existing methods.

CCS CONCEPTS

• **Computing methodologies** → *Natural language processing*; • **Information systems** → *Clustering and classification*; • **Human-centered computing** → *Collaborative and social computing*.

KEYWORDS

cross-domain datasets, neural networks, linguistic bias detection, text tagging, adaptive pre-training, transformer-based models

ACM Reference Format:
Karthic Madanagopal and James Caverlee. 2022. Improving Linguistic Bias Detection in Wikipedia using Cross-Domain Adaptive Pre-Training. In *Companion Proceedings of the Web Conference 2022 (WWW '22 Companion), April 25–29, 2022, Virtual Event, Lyon, France.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/XXXXXX.XXXXXX

1 INTRODUCTION

Wikipedia is a critical platform for organizing and disseminating knowledge, with over 6 million articles shaped by 133 thousand active editors [42]. Furthermore, Wikipedia is heavily relied upon by search engines and other knowledge bases that depend on its quality [22]. To create a reliable encyclopedia, the Wikipedia community is guided by three core policies: (i) always present the content with a "neutral point of view" (NPOV); (ii) all the facts presented in the article need to have verifiable sources with proper attribution; and (iii) no original research [40].

Of the three, NPOV poses key challenges to Wikipedia contributors and editors. For controversial topics (such as politics and current events) it can be difficult to enforce a neutral point of view since some of the information presented is controversial, subjective, and unverifiable. Furthermore, editors may knowingly or unknowingly create bias through their decisions in shaping an article.

With the scale of Wikipedia, the rapidity of edits (about 1.8 edits per second), and the laborious task of resolving NPOV concerns, it is a key challenge to automatically identify biased statements from across Wikipedia. Previous efforts have mainly focused either on (i) manually constructing bias lexicons to identify common linguistic cues (e.g., hedges, weasel words) or (ii) solely focusing on Wikipedia itself as a source of training data for machine learning models [5, 14, 23, 27, 28]. The first approach provides guidance for identifying bias, but may not be appropriate for ongoing detection in dynamic environments like Wikipedia. The second approach relies on a training corpus derived from Wikipedia's NPOV-edits (a collection of edits made to pages that have been tagged as potentially NPOV). However, not all variants of bias may be captured using a Wikipedia-centric NPOV corpus, resulting in models that may miss many instances of biased statements. Indeed, we find that a simple BERT-based bias classifier using this NPOV corpus results in an accuracy of only 73%, with a majority of errors arising in articles on language, literature, politics, and government.

Hence, we investigate in this paper the potential of a *cross-domain pre-training approach* to learn evidence of biased statements from multiple different sources that may provide deeper insights into the kinds of subtle bias that occur on Wikipedia. This approach has two key features: (1) First, we propose to improve the detection of biased statements by leveraging annotated datasets from other domains that are rich in subjectivity and apply recent deep transformer models like BERT in order to more robustly model Wikipedia statements. The proposed approach incorporates evidence from Wikipedia itself, as well as the MPQA Opinion Corpus (which contains news articles annotated for beliefs, emotion, sentiments, etc.) and the Ideological Book Corpus (IBC) (which contains ideologically labeled sentences from U.S. presidential candidates). Since Wikipedia encompasses topics across many domains, we contend that the proposed cross-domain bias detection approach using labeled data from these multiple perspectives could significantly improve the quality of biased statement detection. (2) Second, we adopt a cross-domain

WWW '22 Companion, April 25–29, 2022, Virtual Event, Lyon, France.

ACM ISBN 978-1-4503-9130-6/22/04…$15.00
https://doi.org/10.1145/XXXXXX.XXXXXX

pre-training method to learn contextual relationships between words in a sentence to detect biased statements. Starting with a pre-trained transformer model, we adapt the model using domain-specific unlabeled data (biased language statements) as in Gururangan et al. [12]. Next, we use the domain-adapted model to train a classifier to learn task-specific structural features for cross-domain bias detection. Together, this cross-domain data augmentation method combined with domain adaptive pre-training learns common latent factors and task-specific latent factors across multiple domains, resulting in effective knowledge transfer and improved performance versus hand-crafted features and deep feature synthesis methods used in previous works [15, 28]. The proposed cross-domain pre-training approach with a RoBERTa model outperforms the state-of-the-art bias detection algorithms with a precision of 89% and an F1 score of 87%. Importantly, the proposed model detects subtle forms of bias with higher accuracy than state-of-the-art methods. Furthermore, this study provides the first comprehensive assessment of using cross-domain training to improve automatic detection of biased statements in resources beyond Wikipedia. Finally, we make available this cross-domain bias detection dataset for other researchers to use.

2 RELATED WORK

In this section, we highlight related work on bias in Wikipedia and biased statement classification.

2.1 Subjective Bias in Wikipedia

Wikipedia articles receive around 1,900 revisions on average, but there are concerns of increasing bias [39, 42]. Changes in editor behaviors and their influences in deciding the absolute level of bias were studied by Das and Lavoie et al. [8]. In this work, we concentrate on bias that is introduced by the use of subjective language in presenting information. Wikipedia also contains articles related to business, products and services [41] and sometimes articles are written as advertisements using promotional tone [5]. These promotional pages often use 'peacock' words or phrases that are used to promote a subject [5]. Recasens et al. identified two important types of bias in Wikipedia: (i) framing bias and (ii) epistemological bias [28].

Framing bias is an explicit form of bias that reveals the authors' stance on a particular topic by the use of one-sided or subjective words. It's mostly seen in argument situations, where the speaker takes one side and expresses an opinion strongly opposing it or supporting it. An example is: "She was fired at the end of that episode, and it was *fantastic*." *Epistemological bias* is an implicit and a subtle form of bias that tends to casts a doubt in the expressed information. Identifying epistemological bias can be extremely difficult because if the expressed doubt in the statement is universally accepted, then the fact is unbiased. An example of epistemological bias is: "He *claimed* to be raising money for leukaemia research."

2.2 Biased Statement Classification

Language-based bias detection methods on Wikipedia and for other domains may be grouped into two categories based on the type of features used: (i) manual feature engineering and (ii) deep feature synthesis.

Manual Feature Engineering. Linguistic features like factive verbs, implictive verbs, assertive verbs, and hedges have shown to be effective in detecting both framing and epistemological bias [3]. Anderka et al. [2] used meta-features such as edit histories, reference links, word counts and structural features to measure the quality of an article. Bhonsale et al. [5] used differences in lexical styles in order to identify articles with promotional tone. Wagner et al. [32] used lexical analysis to identify gender bias in Wikipedia articles. Recasens et al. [28] analyzed edit histories of all Wikipedia articles that are disputed for NPOV issues and created a sentence-level bias detection dataset. Then a logistic regression model was developed to detect biased statements by using boolean features that were generated from eight pre-compiled word lists including features like factive verbs, assertive verbs, and implicatve verbs. Entailments and subjectives were found to have higher discriminative power in classifying biased words in a statement. Not all words captured in the bias lexicon are necessarily related to introducing bias, though. We have identified several noise words such as *university* and *marriage*. Roy et. al [29] used frame indicators to identify biased perspectives for different topics. In addition to a lexicon, they used pointwise mutual information for a word in a frame to effectively identify political perspectives. Our proposed approach complements these manual feature engineering approach by learning contextual relationships between words in a sentence to improve bias detection accuracy across different domains through a data augmentation and cross-domain pre-training approach.

Deep Feature Synthesis. Arkajyoti et al. [23] developed a deep learning model to detect political bias using the Ideological Book Corpus, assembled from US congressional floor debate transcripts from 2005. Hube et al. [14] proposed a semi-automated approach to detect statements in Wikipedia with NPOV issues. The first step is the manual construction of a bias lexicon and the second step is to train a machine learning model to detect biased statements. Along with a manually constructed bias lexicon, additional features such as Part of Speech tags (POS) and Linguistic Inquiry Word Count (LIWC) were also used as features for supervised learning. Hube et al. [15] subsequently developed a recurrent neural network based model to detect biased statements in Wikipedia. This is a fully automated approach where a rich representation of the features were automatically constructed using deep learning models. Li et al. [18] adapted pretrained representations to incorporate social and linguistic information to identify political perspectives in news articles. In contrast, we explore the value of leveraging subjective expressions from other domains to improve the performance of Wikipedia bias detection. Also our continual pre-training approach efficiently learns common latent factors and domain specific latent factors across multiple domains, resulting in effective knowledge transfer and improved performance in cross-domain bias detection.

3 BIAS DETECTION METHOD

In this section, we propose to improve Wikipedia bias detection through a combination of domain-adaptive pre-training and task-specific-training that leverages labeled and unlabeled data from multiple related domains.

Preliminaries. As a first step, we trained a simple BERT based bias classifier using the standard labeled NPOV dataset from Wikipedia.

The bias classifier was able to identify biased statements with an accuracy of 73% and F-1 score of 75%. We analysed all of the biased statements that were misclassified and grouped them according to the Wikimedia Foundation's categorization models [4]. A majority of the misclassified statements come from the following three categories (i) Language & Literature (43%), (ii) Politics & Government (26%), and (iii) Sports (22%). This suggests that bias indicators in the NPOV dataset have only a limited coverage of bias types that can arise. Our hypothesis is that leveraging a wider range of subjectivity rich collections could potentially improve the performance of Wikipedia bias detection, as well as provide the foundation for bias detection in other domains.

1. Cross-Domain Indicators of Bias. First, we aim to construct a cross-domain dataset with a wide coverage of biased and unbiased statements. Previous work mainly relies on Wikipedia's NPOV-edits as the main source of training data for bias classifiers [5, 14, 15, 28]. Since NPOV-edits in Wikipedia are necessarily limited in scope and may not capture a variety of common subjectivity expressions, the performance of models that are trained on NPOV-edits have historically been limited (<70% accuracy) and do not adapt well for evolving domains such as Wikipedia [15, 28] (See Table 4). Can a cross-domain dataset overcome these limitations?

2. Additional Pre-Training for Bias Detection. Second, we propose to use supervised domain adaptation techniques that aim to adapt a model trained in a source domain to a new target domain [6, 33] for improved bias statement classification. Through the use of a cross-domain pre-trained RoBERTa model, we show how to improve the performance of cross-domain bias detection model. Many previous methods have relied on pre-BERT machine learning based models. Do we find improvements for biased statement classification along the lines of those in other areas where BERT-based models have been used?

3.1 Data Augmentation

Our first priority is to create a training corpus that is rich in capturing many variants of subjective bias. We studied various research related to bias detection in other domains and selected three sources for assembling our cross-domain training dataset. Wikipedia being our target domain, we used NPOV-edits to extract sentences that capture language patterns related to common writing styles and expressions that impose subjective views. In order to expand our coverage for domain independent expressions related to judgements, interpretations, evaluations and opinions, the MPQA Opinion dataset was selected. Since politics is one of the most biased Wikipedia categories [11], we used the Ideological Book Corpus that is rich in political ideology related language. With MPQA representing a domain independent bias dataset and IBC representing a domain specific dataset (politics), we will be able to better understand the relevance of our approach and also identify what kinds of external datasets could help in improving cross-domain bias detection.

The Wikipedia NPOV Corpus (D_{NPOV}): Our first source is based on Wikipedia itself. We started with the NPOV corpus that was constructed by [28] and then augmented it with new NPOV sentences extracted from recent Wikipedia articles [36]. From the original NPOV corpus, we extracted sentences that had NPOV or peacock tags in their content before the edit.

Then we filtered sentences that contain a minimum of 5 character edits during the revision. Since the original NPOV corpus was created in 2013, in order to capture the new editing style of Wikipedians, we also extracted NPOV sentences from Wikipedia articles that were published between 2013-2019 [36]. Following [28], we analyzed the revision histories for each article and downloaded sentences that were argued for NPOV issues. Sentences that had NPOV tags before the revision were considered as biased sentences and the sentences whose NPOV tags were removed after edits were considered as unbiased sentences. We ignored revisions that were related to missing references, misspellings and punctuation. We also downloaded additional Wikipedia sentences from articles that were tagged as good Wikipedia articles by Wikipedia authors or administrators [37] A total of 107,565 sentences were extracted for our study. The NPOV corpus will help our bias detection model to learn common patterns that are used by Wikipedia editors for imposing subjective views. Some examples of biased statements extracted from the NPOV corpus are:

- *The couple had only been together since early 2014.*
- *It is often labeled a conservative organization.*

MPQA Opinion Corpus (D_{MPQA}): Our second source is based on the MPQA Opinion Corpus. MPQA contains news articles from a wide variety of news sources manually annotated for opinions and other private states (e.g., beliefs, emotions, sentiments, speculations) [35]. The corpus was collected and annotated as part of the summer 2002 NRRC Workshop on Multi-Perspective Question Answering (MPQA). It contains 10,657 sentences collected from 535 documents. A major portion of the articles we identified in the NPOV category are event pages that are written in an opinionated fashion and belong to politics, culture and literature. Since the MPQA corpus is created to capture opinions that are expressed in news related sources, it would be useful to detect bias in event related sentences. Some examples of biased statements extracted from the MPQA corpus are:

- *China criticized the U.S. report's criticism of China's human rights record.*
- *The U.S. fears a spill-over.*

Ideological Book Corpus (D_{IBC}): Finally, we adopt the Ideological Book Corpus (IBC), which contains sentences extracted from speeches of U.S. Presidential candidates [30]. A total of 4,062 sentences were extracted from 112 books and 10 magazine titles and manually labeled with one of three ideological categories: liberal (2,025 sentences), conservative (1,701 sentences) and neutral (600 sentences). The language used in IBC for expressing political ideologies couple help to learn more about adapting our Wikipedia bias detection model for ideological bias. Examples of biased statements from IBC include:

- *An entertainer once said a sucker is born every minute, and surely this is the case with those who support nationalized health care.*
- *They dubbed it the "death tax" and created a big lie about its adverse effects on small businesses.*

Table 1 provides a summary of the biased and unbiased sentences extracted from each corpus.

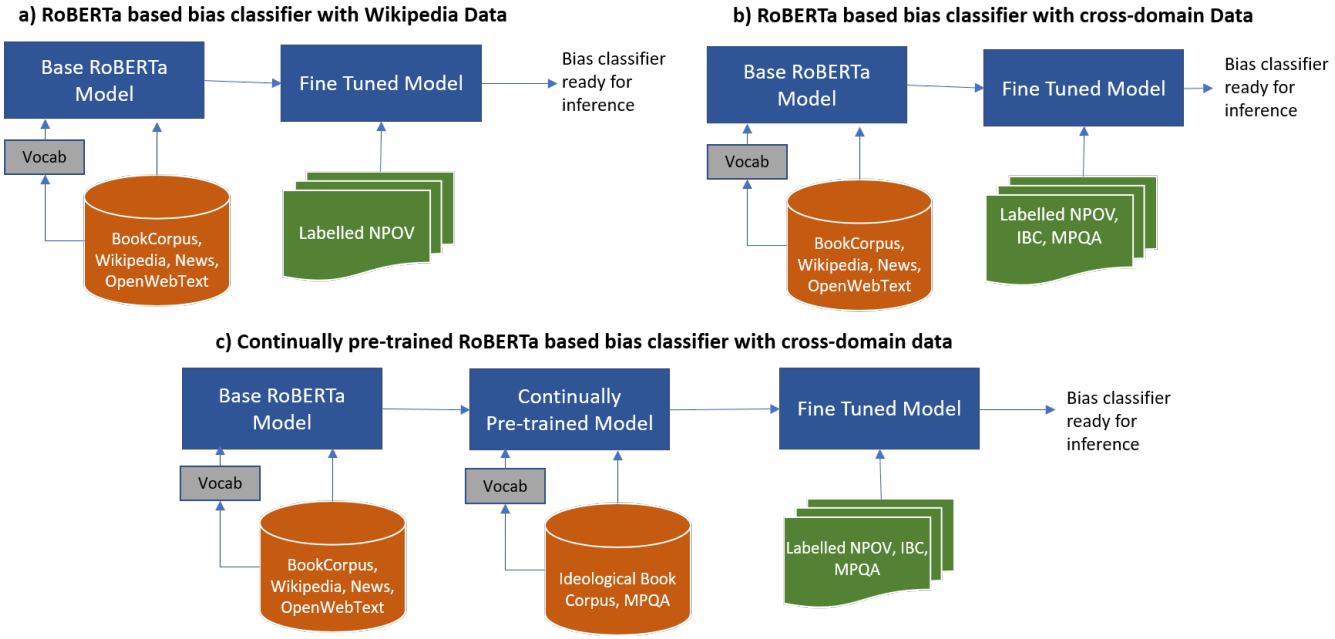

Figure 1: Three different approaches were explored to improve linguistic bias detection in Wikipedia. (a) a RoBERTa based classifier is trained using labeled Wikipedia bias dataset, (b) a RoBERTa based classifier is trained with labeled bias data from multiple domains, and (c) a RoBERTa model is continually pre-trained using unlabeled cross domain data and then fine-tuned to classify biased statements in Wikipedia using labeled datasets from multiple domains like IBC and MPQA.

Corpus	Biased Sentences	Unbiased Sentences	Total
Wikipedia (D_{NPOV})	32,541	75,024	107,565
Opinion Corpus (D_{MPQA})	8,575	42,282	50,857
Ideological Book Corpus (D_{IBC})	3,726	600	4,062

Table 1: Summary of our labeled cross-domain dataset used for building cross-domain bias classifier

3.2 Cross-domain Adaptive Pre-training

The main objective of this task is to design a bias classification model that will take a sentence as an input and produces output that indicates whether the sentence is biased or not. Recent studies on building bias classifiers using neural network models show significant improvement in classification accuracy compared to traditional approaches [15, 27].

One important challenge beyond the use of a bias lexicon is in identifying subtle forms of epistemological bias. Pre-trained contextualized language models have proven to be efficient in incorporating sentence semantics in performing text classifications. Pre-trained language models are domain independent and work well on generic NLP tasks, but continually pre-training and adapting it to the target domain could lead to better performance.

Hence, we adopt a cross-domain pre-trained transformer based deep neural network architecture to capture both the local and global features such as phrases and sentence semantics. It consists of two layers, namely (i) BERT Embedding layer, and (ii) Classification layer.

BERT Embedding Layer: The word embedding layer helps to represent words of a sentence that can capture the semantics of the word depending on their usage in the training corpus. Various pre-trained language models are available for word embeddings such as word2vec, Glove, ELMo and BERT. In our experiments, we have used the base BERT model and the optimized RoBERTa model for our evaluation. RoBERTa is an enhanced BERT model that drops BERT's next-sentence prediction and relies on pre-training with larger batches of data [19].

Classification Layer: The output layer computes a probability of assigning an output label to the input sentence, which is achieved by putting an softmax layer at the end of the BERT embedding layer. The softmax function turns the vector of scores calculated by the BERT layer into a probability distribution.

Pre-trained language models like BERT and RoBERTa are trained on exceptionally large and diverse datasets, contributing to their ability to learn contextual word representations that act as high-quality language features for many downstream NLP tasks. But, lack of domain specific vocabulary in pre-trained models leads to relatively sub-optimal results. Recent research has demonstrated how

additional pre-training on a target domain can improve performance on a target task [1, 7, 12, 31]. Hence, we continually pre-trained the BERT and RoBERTa language models with cross-domain datasets to improve bias classification accuracy across various domains. We adapted a method similar to Gururangan et al. [12], to first perform domain adaptation of a pre-trained model and then performed task-specific training. We evaluated the significance of our cross-domain pre-training technique with different combinations of cross-domain datasets. All our models are implemented using PyTorch [25] and the pre-trained BERT models were downloaded from HuggingFace [43]. We used Nvidia Tesla V100 GPU for continual pre-training of BERT and RoBERTa models; optimization and hyperparameters choices obtained from [9, 20].

4 EVALUATION

In this section, we evaluate the proposed cross-domain pre-training approach by comparing its performance against existing bias detection methods and over various domains (including Wikipedia). Our evaluation is organized around the following research questions:

RQ1: What is the significance of additional cross-domain pre-training in detecting biased statement, with respect to the baseline models?

RQ2: What kind of cross-domain datasets are valuable in accurately identifying subjective bias in Wikipedia?

RQ3: What kind of BERT fine-tuning will help in improving the performance of cross-domain bias detection?

4.1 Evaluation Dataset

The assembled cross-domain training corpus consists of 162,484 sentences with a split of 107,565 Sentences from Wikipedia, 50,857 sentences from MPQA and 4,062 sentences from NPOV dataset. From the total sentences available, we used a 70/30 split for our training and testing. For both datasets, we made sure they have the same percentage of biased (27%) and unbiased statements (73%).

4.2 Evaluation Metrics

Primarily we want our model to be highly accurate in identifying biased statements. At the same time, we also want our model to capture as many subjective statements without compromising its classification accuracy. Hence we selected precision, recall, and F1-score as our metrics.

4.3 Baselines

To compare the performance of the proposed model against existing bias detection models, we developed two baselines:

Baseline-1: A bag of words based text classifier that uses a curated set of bias lexicons collected from multiple subjectivity based studies [13, 16, 17, 28, 34].

Baseline-2: A logistic regression model that uses a set of manually-curated 32 linguistic features such as factive verbs, implicatives, hedges and subjective intensifiers prescribed in [28].

Baseline-1 allows us to compare against a simple bag-of-words model using bias lexicons. Baseline-2 allows us to compare against a traditional machine learning model developed with extensive feature engineering.

4.4 Model Selection

Before our experimental evaluation of cross-domain biased statement classification, we first investigate the design of our BERT-based model. The two main objectives for this task are (i) identify which of our deep learning architectures are best for the training and evaluation of the cross-domain bias classifier, and (ii) create baseline results for comparing the performance of our models trained on Wikipedia alone vs. cross-domain resources. We adopt the Wikipedia NPOV dataset as our train-test dataset and then train three different models (i) GloVe based classifier (CLS_{GloVe}), (ii) BERT based classifier($CLS_{BERT(Base)}$), (iii) RoBERTa based classifier($CLS_{RoBERTa(Base)}$). The pre-trained GloVe, BERT and RoBERTa models used in this experiment were downloaded from the Stanford and HuggingFace repositories respectively [26, 43]. The BERT based classifier used a pre-trained BERT model with a single linear layer added to the top in order to perform sentence classification. In this task, we trained and tested all the models using the Wikipedia NPOV corpus.

Model	Precision	Recall	F1-score
Baseline-1	56.24	**86.74**	68.24
Baseline-2	69.42	63.74	66.47
CLS_{GloVe}	69.81	69.81	70.08
$CLS_{BERT(Base)}$	73.87	71.26	72.54
$CLS_{RoBERTa\ (Base)}$	**77.92**	76.24	**77.07**

Table 2: The $CLS_{RoBERTa\ (Base)}$ showed significant improvement in classifying biased statement on Wikipedia NPOV corpus. Bold indicates best results.

The detailed results for the three deep models compared with the baseline models are shown in Table 2. The bias lexicon based model (Baseline-1) had an accuracy of 56% with a recall of 86%. The lexicon based model favored classifying sentences as biased as seen by its high recall and low precision. The linguistic based model (Baseline-2) provided balanced results with 69% accuracy and 63% recall. The GloVe based classifier had similar accuracy as Baseline-2 but it's recall was higher. The BERT based classifier yielded a classification accuracy of 75% with a recall of of 73%. The RoBERTa based model had the best performance with 77% accuracy and 76% recall. From manual verification of the results, we observed that the RoBERTa model is able to perform better sense disambiguation of the bias lexicons in identifying biased sentences relative to other models. To illustrate, the following sentences were correctly classified by the RoBERTa based model.

- *Biased: Since 1999, at least 148 people have died in the United States and Canada after being shocked with Tasers by police officers, according to the ACLU.*
- *Unbiased: The ACLU alleges that, since 1999, at least 148 people have died in the United States and Canada after being shocked with Tasers by police officers.*

Even though *shocked* is a subjective word, the BERT and RoBERTa model is able to learn the use of *alleges* to classify the sentences correctly. The recall of the classifier also showed improvement in

Model	Training Corpus	Precision	Recall	F1-score
$CLS_{RoBERTa(Base)}$	$[D_{NPOV}]$	77.92	76.24	77.07
$CLS_{RoBERTa(NPOV + MPQA + IBC)}$	$[D_{NPOV}]$	80.92	79.57	80.24
$CLS_{RoBERTa(NPOV + MPQA + IBC)}$	$[D_{NPOV}] + [D_{MPQA}]$	86.39	83.54	84.94
$CLS_{RoBERTa(NPOV + MPQA + IBC)}$	$[D_{NPOV}] + [D_{MPQA}] + [D_{IBC}]$	**89.41**	**85.94**	**87.64**

Table 3: The impact of cross-domain pre-training: adding additional sources of biased statements (MPQA and IBC) improves precision, recall, and F1-score. Domain data used to continually pre-train RoBERTa model is shown in bracket for pre-trained models. Bold indicates best results.

Testing Domain	Baseline			$RoBERTa_{NPOV + MPQA + IBC}$		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Wikipedia (NPOV)	69.42	63.74	66.46	89.41	85.94	87.64
Opinion Corpus (MPQA)	79.00	76.00	77.50	91.02	91.90	91.46
Ideological Book Corpus (IBC)	63.00	68.21	65.50	69.97	83.46	76.12

Table 4: Performance comparison of the baseline and cross-domain model on different test domains.

BERT and RoBERTa models due to its ability to capture different variants of bias inducing lexicons. In the model selection experiment, the pre-trained BERT and RoBERTa models downloaded from [43] were used as-is. More detailed experiments on the fine tuned RoBERTa model and other parameters will be presented in the subsequent sections.

4.5 Impact of cross-domain pre-training

To demonstrate the value of cross-domain pre-training, we trained the best model (RoBERTa based classifier) using three different combinations of our cross-domain training corpus. The detailed results of this evaluation are presented in Table 3. We note that the two vanilla models ($BERT_{Base}$ and $RoBERTa_{Base}$) that were trained over the Wikipedia dataset (D_{NPOV}) alone outperformed the baselines by a margin of 4-8% in classification accuracy. This shows that transformer-based models are powerful enough to learn domain specific sentence structures that are relevant to detecting language induced subjective bias. In the next task, instead of directly fine-tuning base pretrained transformer models (RoBERTa) we first continually pre-trained them with domain specific data and then fine-tuned for the bias detection task. The continual pre-training of base RoBERTa model helps to improve the performance by converting its context independent vectors into context sensitive vectors. In the next task, we continually pre-trained RoBERTa model with cross-domain datasets before training our bias classifier and evaluated its performance. For continual pre-training of transformer models, we used source documents that were used to create our labelled cross-domain dataset. Since IBC has a very small volume of source documents, news articles were used from the CC-News-En dataset [21], the HuffPost dataset [24] and the BBC dataset [10]. Around 600,000 articles were used for the continual pre-training process.

The four models achieved strong performance on all three co-training datasets, especially for $D_{NPOV} + D_{MPQA} + D_{IBC}$, in which our model has a 12% margin over the best previous result. There is a 6% increase in classification accuracy between our models trained with D_{NPOV} and $D_{NPOV} + D_{MPQA}$, which confirms our initial hypothesis that using biased statements from other domains will help to improve bias detection in Wikipedia. Although we have assumed that the introduction of additional domains beyond MPQA will further improve the performance of our bias classifier, only a small margin (2-3%) improvement is observed between $D_{NPOV} + D_{MPQA}$ and $D_{NPOV} + D_{MPQA} + D_{IBC}$. We identify two reasons behind this observed small improvement: (i) MPQA captures a majority of subjective expressions and IBC's addition only helped to augment political language patterns to the mix, and (ii) limited amount of domain knowledge from such a small IBC corpus.

From the observed results, we learned that the performance of cross-domain models depends on the amount of knowledge overlap between the source and target domain. Since NPOV and MPQA contain generic subjective bias statements, their combined model performed better. In the other case, the overlap between NPOV and IBC is limited in terms of politically biased statements and the performance improvement is not significant. Additionally, our cross-domain pre-trained model also provided a small margin of improvement in bias classification accuracy. Here are few examples of the biased sentences that were identified:

- *Spread over 40-acre campus, AKGEC has excellent infrastructure with well-planned complexes.*
- *Obama's most significant legacy is generally considered to be the Patient Protection and Affordable Care Act (PPACA).*
- *The arrangements are sublime, the performances infallible.*
- *Privacy remains an ongoing problem for Facebook.*
- *It will be a shock to most of the people, but homeopathy is successful in curing osteoporosis with the help of two main remedies, aurum met.*
- *He also favors drastic reductions in government spending and the elimination of corporate welfare.*

Pre-trained Model	Training Corpus	Precision	Recall	F1-score
$\text{RoBERTa}_{\text{Base}}$	$[D_{NPOV}]$	77.63	76.24	76.92
$\text{RoBERTa}_{\text{Base}}$	$[D_{NPOV}]$ + $[D_{MPQA}]$	**83.62**	**82.95**	**83.62**
$\text{RoBERTa}_{\text{Base}}$	$[D_{NPOV}]$ + $[D_{IBC}]$	82.73	81.67	82.20
$\text{RoBERTa}_{\text{Base}}$	$[D_{NPOV}]$ + $[D_{MPQA}]$ + $[D_{IBC}]$	83.97	81.17	82.55
$\text{RoBERTa}_{\text{NPOV + MPQA + IBC}}$	$[D_{NPOV}]$	80.92	79.57	80.24
$\text{RoBERTa}_{\text{NPOV + MPQA + IBC}}$	$[D_{NPOV}]$ + $[D_{MPQA}]$	86.39	83.54	84.94
$\text{RoBERTa}_{\text{NPOV + MPQA + IBC}}$	$[D_{NPOV}]$ + $[D_{IBC}]$	84.11	82.76	83.43
$\text{RoBERTa}_{\text{NPOV + MPQA + IBC}}$	$[D_{NPOV}]$ + $[D_{MPQA}]$ + $[D_{IBC}]$	**89.41**	**85.94**	**87.64**

Table 5: The ablation study shows cross-domain pre-training improved precision, recall and f1-score on RoBERTa based classifiers. Since the Wikipedia is the target domain, $[D_{NPOV}]$ is used in all training corpus. Continually pre-trained RoBERTa based classifier provided best results. Bold indicates best results.

4.6 Comparison across various domains

Since our cross-domain training corpus has 66% of sentences related to Wikipedia, it performed well on our Wikipedia test dataset. The performance improvement is mainly attributed to 34% of sentences coming from other domains. To understand the performance of our bias classifier on other domains, we tested the classifier on test datasets created using the MPQA opinion corpus and the IBC corpus. The sentences used for this testing are not used for training the model. Detailed results of our evaluation of various domain related test datasets are presented in Table 4. We observed a 12% improvement in classification accuracy on the MPQA opinion corpus over OpinionFinder 2.0. This result might be due to the significant overlap in the subjective language between the two domains (Wikipedia and News articles). Interestingly, the performance of our model on the IBC test dataset is not significant. Our reasoning behind this performance gap is due to the overloaded political language on IBC and the relatively fewer number of politically unbiased sentences (<5%) in our NPOV corpus. Also, the presence of presumptive bias in political statements can be difficult to classify given the small percentage of political statements in our cross-domain bias corpus. By balancing the mix of sentences in our cross-domain corpus, we hope to improve the performance of our bias classifier in IBC-like domains as well.

4.7 Ablation Study

To understand the significance of the proposed cross-domain pre-training approach, we designed this ablation task to compare bias classification using two different models (i) a classifier trained on Base RoBERTa model, and (ii) a classifier trained on cross-domain pre-trained RoBERTa model. Except the usage of different pre-trained models, the training stage and data used are same across all the models. Instead of training RoBERTa models from scratch, we fine-tuned them using our cross-domain datasets. Here we compared our best model with base and cross-domain pre-trained models and the results of our evaluation are presented in Table 5.

Augmenting labeled data from other domains in the training phase shows improvement in all the models relative to baselines. In case of classifiers trained with base pre-trained models, the addition of the MPQA dataset improved the recall by 5%, but we did not observe a significant change in performance after adding the IBC dataset. In contrast, the cross-domain pre-trained models showed consistent 3-4% improvement in bias classification accuracy. This experiment confirms that the domain adapted pre-trained models through cross-domain pre-training provide more discriminative power and perform effective knowledge transfer across different domains in classifying biased sentences in comparison with non-domain adapted pre-trained models.

4.8 Case Study: Gun Control

Our evaluation so far has focused on labeled datasets for which we have ground truth. In this last experiment, we collected a set of unlabeled sentences from one of the most controversial and discussed articles in Wikipedia: Gun Control [38]. Following an approach similar to Hube et al., we collected 135 sentences by analyzing the revision histories and filtering them by NPOV issues [15]. Then we manually reviewed the 135 sentences and identified 32 of them as biased statements. Our bias detection model classified 38 sentences as biased, of which 28 of them are true positives and 10 are false positives, giving us an accuracy of 73% and a recall of 87%. These results further confirm potential benefits of cross-domain continual pre-training. Here are few example biased sentences identified by our model.

- *Opponents of gun control sometimes argue that wide legal ownership of pistols, including the right to carry them concealed, actually deters crime rather than increases it.*
- *The three shared anti-government views, including opposition to gun control and anger at the federal government's handling of the Waco Siege and the incident at Ruby Ridge.*

The first sentence was written with epistemological bias that implicitly casts a doubt in the expressed information by using *sometimes*. The framing of the second sentence reveals the authors' stance on the topic by coining the expressed information as an *anti-government* view.

5 CONCLUSION

In this work, we investigated the value of leveraging knowledge from other subjectivity rich domains for enhancing the performance of Wikipedia bias detection. Specifically, we constructed a cross-domain training corpus and designed a textcross-domain adaptive pre-training based model to classify a sentence in Wikipedia as being biased or not. In contrast to most previous methods that use training datasets from Wikipedia articles such as NPOV-edits, we show that the proposed approach detects biased statements in Wikipedia more accurately than existing state-of-the-art models by leveraging a rich pre-trained language model and fine-tuning it with a cross-domain training corpus. Additionally, we showed that a cross-domain trained model also performs well in detecting biased statements in other domains such as news articles and political speeches.

REFERENCES

[1] Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. *arXiv preprint arXiv:1904.03323* (2019).

[2] Maik Anderka, Benno Stein, and Nedim Lipka. 2012. Predicting quality flaws in user-generated content: the case of wikipedia. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. 981–990.

[3] CJ Hutto Dennis Folds Scott Appling. 2017. Computationally Detecting and Quantifying the Degree of Bias in Sentence-Level Text of News Stories. (2017).

[4] Sumit Asthana and Aaron Halfaker. 2018. With few eyes, all hoaxes are deep. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–18.

[5] Shruti Bhosale, Heath Vinicombe, and Raymond Mooney. 2013. Detecting promotional content in wikipedia. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. 1851–1857.

[6] Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*. 92–100.

[7] Tuhin Chakrabarty, Christopher Hidey, and Kathleen McKeown. 2019. IMHO fine-tuning improves claim detection. *arXiv preprint arXiv:1905.07000* (2019).

[8] Sanmay Das, Allen Lavoie, and Malik Magdon-Ismail. 2016. Manipulation among the arbiters of collective intelligence: How Wikipedia administrators mold public opinion. *ACM Transactions on the Web (TWEB)* 10, 4 (2016), 1–25.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[10] Derek Greene and Padraig Cunningham. 2006. Practical solutions to the problem of diagonal dominance in kernel document clustering. In *Proceedings of the 23rd international conference on Machine learning*. 377–384.

[11] Shane Greenstein and Feng Zhu. 2012. Is Wikipedia Biased? *American Economic Review* 102, 3 (2012), 343–48.

[12] Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. *arXiv preprint arXiv:2004.10964* (2020).

[13] Joan B Hooper. 1975. On assertive predicates. In *Syntax and Semantics volume 4*. Brill, 91–124.

[14] Christoph Hube and Besnik Fetahu. 2018. Detecting biased statements in wikipedia. In *Companion Proceedings of the The Web Conference 2018*. 1779–1786.

[15] Christoph Hube and Besnik Fetahu. 2019. Neural based statement classification for biased language. In *Proceedings of the twelfth ACM international conference on web search and data mining*. 195–203.

[16] Ken Hyland. 2018. *Metadiscourse: Exploring interaction in writing*. Bloomsbury Publishing.

[17] Lauri Karttunen. 1971. Implicative verbs. *Language* (1971), 340–358.

[18] Chang Li and Dan Goldwasser. 2021. Using social and linguistic information to adapt pretrained representations for political perspective identification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 4569–4579.

[19] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[20] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR* abs/1907.11692 (2019). arXiv:1907.11692 http://arxiv.org/abs/1907.11692

[21] Joel Mackenzie, Rodger Benham, Matthias Petri, Johanne R Trippas, J Shane Culpepper, and Alistair Moffat. 2020. CC-News-En: A large English news corpus. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 3077–3084.

[22] Connor McMahon, Isaac Johnson, and Brent Hecht. 2017. The substantial interdependence of Wikipedia and Google: A case study on the relationship between peer production communities and information technologies. In *Eleventh international AAAI conference on web and social media*.

[23] Arkajyoti Misra and Sanjib Basak. 2016. Political bias analysis.

[24] Rishabh Misra and Jigyasa Grover. 2021. *Sculpting Data for ML: The first act of Machine Learning*.

[25] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alche-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[26] 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)*. 1532–1543.

[27] Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2019. Automatically Neutralizing Subjective Bias in Text. *arXiv preprint arXiv:1911.09709* (2019).

[28] Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1650–1659.

[29] Shamik Roy and Dan Goldwasser. 2020. Weakly supervised learning of nuanced frames for analyzing polarization in news media. *arXiv preprint arXiv:2009.09609* (2020).

[30] Yanchuan Sim, Brice DL 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*. 91–101.

[31] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification?. In *China National Conference on Chinese Computational Linguistics*. Springer, 194–206.

[32] Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. 2015. It's a man's Wikipedia? Assessing gender inequality in an online encyclopedia. In *Ninth international AAAI conference on web and social media*.

[33] Xiaojun Wan. 2009. Co-training for cross-lingual sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-volume 1*. Association for Computational Linguistics, 235–243.

[34] Janyce Wiebe and Ellen Riloff. 2005. Creating subjective and objective sentence classifiers from unannotated texts. In *International conference on intelligent text processing and computational linguistics*. Springer, 486–497.

[35] Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation* 39, 2-3 (2005), 165–210.

[36] Wikipedia. 2020. Category:All NPOV disputes — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?oldid=949739383 [Online; accessed 09-April-2020].

[37] Wikipedia. 2020. Category:Lists of good articles — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?oldid=906447739 [Online; accessed 06-May-2020].

[38] Wikipedia. 2020. Gun control — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?oldid=954608308 [Online; accessed 15-May-2020].

[39] Wikipedia. 2021. Criticism of Wikipedia — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?oldid=948675457 [Online; accessed 08-April-2021].

[40] Wikipedia. 2021. Wikipedia:Neutral point of view — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?oldid=949318052 [Online; accessed 07-April-2021].

[41] Wikipedia. 2021. Wikipedia:Spam — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?oldid=937664745 [Online; accessed 08-November-2021].

[42] Wikipedia. 2021. Wikipedia:Statistics — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Wikipedia:Statistics [Online; accessed 07-April-2021].

[43] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2020. Pretrained BERT Models. https://huggingface.co/models