Towards Ongoing Detection of Linguistic Bias on Wikipedia

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 critical platform for organizing and disseminating knowledge. One of the key principles of Wikipedia is *neutral point of view (NPOV)*, so that bias is not injected into objective treatment of subject matter. As part of our research vision to develop resilient bias detection models that can self-adapt over time, we present in this paper our initial investigation of the potential of a cross-domain transfer learning approach to improve Wikipedia bias detection. The ultimate goal is to future-proof Wikipedia in the face of dynamic, evolving kinds of linguistic bias and adversarial manipulations intended to evade NPOV issues. We highlight the impact of incorporating evidence of bias from other subjectivity rich domains into further pre-training a BERT-based model, resulting in strong performance in comparison with traditional methods.

KEYWORDS

Language Bias, Wikipedia Quality, NPOV

ACM Reference Format:
Karthic Madanagopal and James Caverlee. 2021. Towards Ongoing Detection of Linguistic Bias on Wikipedia. In *Proceedings of the Web Conference 2021 (WWW '21), April 19–23, 2021, Ljubljana, Slovenia.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3442442.3452353

1 INTRODUCTION

Wikipedia is one of the most popular open-source encyclopedia that is heavily relied upon by search engines and other knowledge bases that rely on the quality of its information. With its crowd and expert produced knowledge, Wikipedia's "neutral point of view" (NPOV) is a core principle that aims towards improving the reliability and quality of articles. NPOV guidelines expect all Wikipedia articles to be written "fairly, proportionately, and as far as possible without editorial bias"[10].

And yet, controversial topics (such as politics and current events) may make it 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 [6]. Indeed, the scale of Wikipedia, the rapidity of edits (about 1.8 edits per second), and the laborious task of resolving NPOV concerns have motivated significant recent research in building tools to automatically identify biased statements from across Wikipedia, e.g., [3, 5, 6]. These and related works 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, 6]. Both assumptions, however, may limit the ability of bias detection models to robustly adapt in the face of a dynamic, evolving resource like Wikipedia. For example, new kinds of linguistic bias could be injected that are under-represented (or entirely missing) from historic training data, and Wikipedia contributors may learn to evade NPOV issues by adversarially manipulating their writing style and other behaviors. Hence, the *overall vision* of this research project is to develop resilient bias detection models that can self-adapt over time. Such models should be robust to changes in editor behaviors and to new subjective writing styles that have never before been seen by the Wikipedia community.

As a first step toward this vision, we report in this paper our initial investigation into the potential of a cross-domain transfer learning approach to improve Wikipedia bias detection. The key idea is to learn both common latent factors of bias and domain-specific latent factors of bias from across multiple domains (beyond just Wikipedia). In this way, evidence of bias may be rapidly incorporated from multiple sources to provide new insights. Concretely, we exploit two additional datasets that are rich in subjectivity: the MPQA Opinion Corpus that contains news articles annotated for beliefs, emotion, sentiments, etc. [9], and the Ideological Book Corpus (IBC) that contains ideologically labeled sentences from U.S. presidential candidates [7]. We explore the potential of transferring evidence of bias from these domains to Wikipedia, where we find a significant improvement over the current state-of-the-art in Wikipedia bias detection. We find strong performance by both the further pre-training of BERT-based models with unlabeled cross-domain datasets and later training the bias detection classifier with labeled cross-domain datasets. Together, these findings demonstrate the potential of incorporating new sources of bias for improving ongoing detection on Wikipedia.

2 METHODS

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

2.1 Baselines

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

- **BoWM:** A bag of words based text classifier that uses a curated set of bias lexicons collected from multiple subjectivity based studies [2, 6].

WWW '21, April 19–23, 2021, Ljubljana, Slovenia

ACM ISBN 978-1-4503-8312-7/21/04.
https://doi.org/10.1145/3442442.3452353

Table 1: Experimental results of cross-domain pre-training on Wikipedia Dataset

Pre-trained Model	Training Corpus	Precision	Recall	F1-score
BoWM	$[D_{NPOV}]$	0.5624	0.8674	0.6824
LRM	$[D_{NPOV}]$	0.6942	0.6374	0.6647
BERT	$[D_{NPOV}]$	0.7387	0.7126	0.7254
RoBERTa	$[D_{NPOV}]$	0.7792	0.7624	0.7707
$RoBERTa_{fine-tuned}$	$[D_{NPOV}]$	0.8092	0.7957	0.8024
$RoBERTa_{fine-tuned}$	$[D_{NPOV}]$ + $[D_{MPQA}]$	0.8639	0.8354	0.8494
$RoBERTa_{fine-tuned}$	$[D_{NPOV}]$ + $[D_{MPQA}]$ + $[D_{IBC}]$	0.8941	0.8594	0.8764

- **LRM:** 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 [6].
- **BERT:** A BERT-based text classifier to detect biased statements [1].
- **RoBERTa:** A variant of BERT that has demonstrated strong performance in many domains [4].

2.2 Proposed Approaches

As a first step, we tested the four baseline classifiers trained over NPOV-edits only (following the style of much previous research). We carefully analysed the biased statements that are misclassified by the baseline classifiers and grouped them by their topic categories. The top three categories are (i) Language & Literature (43%), (ii) Politics & Government (26%), and (iii) Sports (22%). These errors suggest the possibility that incorporating additional sources of bias could improve coverage of the kinds of biased statements made on Wikipedia.

- **Data augmentation:** Hence, we first construct a cross-domain dataset with a wide coverage of biased and unbiased statements. Based on our initial experiments, we adopted Wikipedia NPOV (D_{NPOV}) plus two additional datasets: (i) the MPQA Opinion Corpus (D_{MPQA}) that contains news articles from a wide variety of news sources manually annotated for opinions and other private states (e.g., beliefs, emotions, sentiments, speculations) [9]; and (ii) the Ideological Book Corpus (IBC) (D_{IBC}) that contains ideologically labeled sentences (covering liberal, conservative, and neutral) from speeches of U.S. presidential candidates [7].
- **Cross-domain Adaptation:** Second, we explore the potential of cross-domain adaptation of BERT-based models for improved bias detection. Recent research has demonstrated how additional pre-training on a target domain can improve performance on a target task [8]. Hence, we further pre-trained a BERT and a RoBERTa-based language model with the cross-domain datasets to explore if evidence from these new sources could improve Wikipedia bias detection.

3 FINDINGS

We report the precision, recall and f1-score across different models in Table 1. In the first experiment, we train models solely with the Wikipedia NPOV dataset (ignoring the additional data augmentation and cross-domain adaptation). Focusing on the top-half of Table 1, we find that the RoBERTa-based model yields the highest f1-score of 77%. Unsurprisingly, these large language models improve upon traditional bias detection methods.

In the second experiment, we fine-tuned a RoBERTa pre-trained model with unlabelled sentences from the NPOV corpus and then trained our bias classifer with labelled sentences from the NPOV corpus. We observed the $RoBERTa_{fine-tuned}$ classifier showed significant improvement (3% increase in f1-score) compared to baseline models. This experiment confirms that additional pre-training helps to adapt the pre-trained model to the target task.

In the final experiment, we fix RoBERTa with fine-tuning as the baseline method (since it performed the best trained solely over Wikipedia NPOV). We then add the MPQA and the IBC datasets. The two RoBERTa-based classifiers trained on cross-domain datasets achieved strong performance in comparison with the baseline models, especially for D_{NPOV} + D_{MPQA} + D_{IBC}. Adding an extra layer to the RoBERTa model and training it with cross-domain datasets provided initial improvements. We observed a significant margin of improvement (12%) by both further pre-training a BERT-based model with the unlabeled cross-domain dataset and later training the classifier with labeled cross-domain dataset (See Table 1).

4 DISCUSSION

These initial experiments demonstrate that cross-domain data augmentation and pre-training can help to build a more robust bias classifier that is able learn linguistic patterns from multiple (non-Wikipedia) domains. These results suggest the potential of such approaches to identify newer and more subtle forms of subjective bias that are emerging in Wikipedia articles. Open questions we are exploring in our continuing research include: Can we also generalize this bias classifier approach to perform well on other target domains beyond Wikipedia? What is the value of additional datasets from other domains for improving the performance of our cross-domain bias classifier? Is it a case of more is always better, or do we need to develop techniques to carefully incorporate these additional sources? What impact do other variants of BERT (e.g., ALBERT, and DistilBERT) have? Finally, a related goal is to develop new tools to automatically rewrite a subjectively biased statement into a neutral form while preserving the fact expressed in the input text.

REFERENCES

[1] 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).

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

[3] 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.

[4] 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).

[5] 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).

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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