

Accenture at CheckThat! 2023: Learning to Detect Political Bias of News Articles and Sources

Sieu Tran¹, Paul Rodrigues¹, Benjamin Strauss¹ and Evan M. Williams²

¹Accenture, 1201 New York Ave NW, Washington, DC 20005, United States

²Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, United States

Abstract

This paper introduces the methodology of Team Accenture for the CLEF CheckThat! shared task on identifying political biases in news articles and news sources. We utilize machine back-translation to augment the minority classes in datasets labeling article and news source bias in three categories-Left, Center, and Right, and used this augmented data to fine-tune RoBERTa transformer models. This was the highest ranking strategy in the shared task for detecting both political bias of a news article (at 0.473 Mean Average Precision) as well as for detecting political bias of a news source (at 0.549 mean average precision).

Keywords

bias detection, political bias, news analysis, data-driven journalism

1. Introduction

Political bias in news articles can jeopardize an article's reliability. Biased articles can employ loaded language, lack proper context, can frame a story selectively, or can include outright falsehoods [1, 2]. Consumption of partisan political outlets has been linked to numerous real-world behavioral differences. Responses to COVID-19 were found by numerous studies to be strongly partisan. In 2020, Gollwitzer et al. found that consumption of the US conservative media outlet, Fox News, was associated with reduced physical distancing and increased vaccine hesitancy [3, 4].

Consequently, political bias detection has become an important task. [1] propose a headline attention network to detect political bias in Telugu newspapers. [5] use Copula Ordinal Regression (COR) models to jointly predict news media reliability and bias. [6] provides an analysis of linguistic features of biased domains. [7] demonstrate that biased news domains form link communities. Other researchers have employed political bias detection methods on non-news datasets, including the political bias of congressional speeches [8], of cable news channels [9], and of YouTube videos [10].

CheckThat! 2023 Task 3A provides bias-labeled articles that teams must classify as 'Left', 'Center', or 'Right'. Task 3B provides bias-labeled news-sites and a collection of articles scraped

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18-21, 2023, Thessaloniki, Greece

✉ sieu.tran@accenturefederal.com (S. Tran); paul.rodrigues@accenturefederal.com (P. Rodrigues);


b.strauss@accenturefederal.com (B. Strauss); emwillia@andrew.cmu.edu (E. M. Williams)

🆔 0000-0003-0017-4329 (S. Tran); 0000-0002-2151-636X (P. Rodrigues); 0000-0002-0224-424X (B. Strauss);

0000-0002-0534-9450 (E. M. Williams)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

from each news site. Given the news source and the articles, the goal is to classify URLs of news outlets as ‘Left’, ‘Center’, or ‘Right’. Teams are evaluated using mean absolute error (MAE) [11] where lower MAE is better. The setup of the task is similar to the article-level fake-news domain detection setup used in [12]. However, Subtask 3B uses domain bias labels rather than domain reliability labels.

In this work, we describe that data augmentation and fine-tuning approach employed by Team Accenture for CheckThat! lab subtasks 3A and 3B. Of the four teams that submitted, the Accenture team achieved the best overall MAE for 3A (0.473). Of the two teams that submitted, the Accenture team achieved the best overall MAE for 3B (0.549).

2. Exploratory Analysis

Table 1 shows the number of samples and unique word counts for each of the datasets provided. We see that while the training set for news article bias consists of a much larger training sample than the news media source bias, it has significantly less number of unique words. We would hypothesize that a larger quantity of unique words would yield models of higher performance.

Table 1
Dataset Descriptions

Task	Modeling Group	# of Source	# of Articles	Unique Words
News Article Bias	Train		45,066	47,054
News Article Bias	Test		5,198	21,503
News Article Bias	Validation		5,008	19,752
News Media Source Bias	Train	817	6,994	89,007
News Media Source Bias	Test	102	896	34,372
News Media Source Bias	Validation	104	878	37,484

2.1. Label Balance

As shown in Figure 1, all of the datasets provided by the CheckThat! organizers had label bias which skewed each dataset towards articles that were labeled class 2 (Right).

2.2. WordPiece Analysis

Transformer models utilize WordPiece tokenization schemes that are dependant on the model being evaluated. At the time of pre-training, the WordPiece algorithm determines which pieces of words will be retained, and which will be discarded. We present our analysis in Table 2. Unexpectedly, the RoBERTa tokenizers we used did not return UNK tokens on any dataset provided by the CLEF CheckThat! organizers.

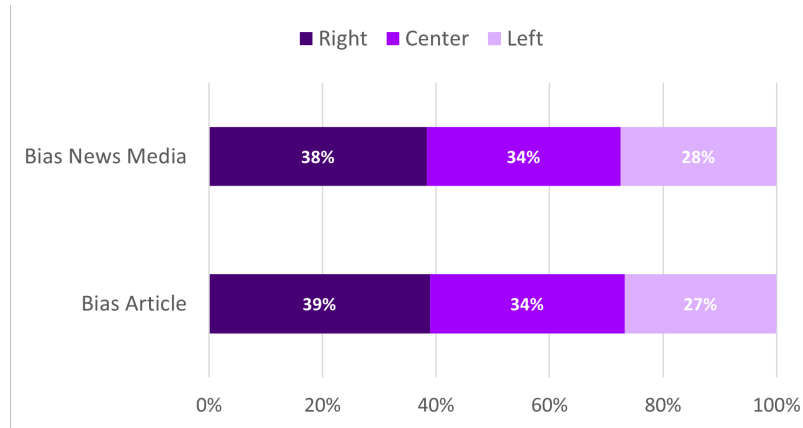


Figure 1: Label distribution across training sets

Table 2

Token Distribution in Data for Each Task.

Task	Tokenizer Type	Modeling Set	WordPiece
News Article Bias	RoBERTa-based	Train	4,336,612
		Test	614,064
		Validation	482,337
News Media Source Bias	RoBERTa-based	Train	7,324,674
		Test	985,660
		Validation	1,128,531

3. Transformer Architectures and Pre-Trained Models

In this work, we utilize RoBERTa models. The Bidirectional Encoder Representation Transformer (BERT) is a transformer-based architecture that was introduced in 2018 [13, 14]. BERT has had a substantial impact on the field of NLP, and achieved state of the art results on 11 NLP benchmarks at the time of its release. RoBERTa, introduced by [15], modified various parts of BERTs training process. These modifications include more training data, more pre-training steps with bigger batches over more data, removing BERT’s Next Sentence Prediction, training on longer sequences, and dynamically changing the masking pattern applied to the training data [16]. For this work, we fine-tune *roberta-large* [17]. The English RoBERTa model contains 50,265 WordPieces.

4. Method

4.1. Data Augmentation

The organizers provided a training and a development set for each language. We use the provided training set and development set to create internal training and validation sets for experimentation. We use the test set provided by organizers as a hold-out test set.

Table 3
Average Sentence BLEU Score for Each Back-translation Scheme

Task	Label Class	Back-translation	Average Sentence BLEU Score
News Article Bias	0 (Left)	EN > ES > EN	0.504
News Article Bias	1 (Center)	EN > ES > EN	0.481
News Media Source Bias	0 (Left)	EN > ES > EN	0.491
News Media Source Bias	1 (Center)	EN > ES > EN	0.517

Table 4
New Tokens in Machine Translated Text

Task	Label Class	Back-translation	Unique tokens in source	Unique tokens in MT	New Tokens in MT
News Article Bias	0 (Left)	EN > ES > EN	61771	55982	19045
News Article Bias	1 (Center)	EN > ES > EN	27017	25498	6916
News Media Source Bias	0 (Left)	EN > ES > EN	36221	31591	11451
News Media Source Bias	1 (Center)	EN > ES > EN	22062	19561	6545

For each article, training data was augmented using back-translation provided by AWS Translate. We appended back-translated left- and center-labeled articles to the training set. In our 2021 experiment [18], we found that this form of augmentation resulted in a significant increase in recall and F1 score for check-worthy tweets. For both article and news source classification, we used Spanish as the pivot language. Due to significant sample imbalance in the training sets for both tasks, we augmented the 0 (Left)- and 1 (Center)-class until the samples are balanced. Specifically, for classification of article bias, we augmented 5,000 samples of the 0-class and 2,000 of the 1 (Center)-class. For classification of news source bias, we augmented 700 samples of the 0 (Left)-class and 300 of the 1 (Left)-class. Table 3 shows the BLEU score for each back-translation scheme. The lower the score, the more divergent the translation to the original text. In a machine translation workflow we would wish to maximize the BLEU score for the best translation. In a data augmentation workflow, we wish to introduce variation to the training data.

4.2. Classification

For the Article Bias Classification RoBERTa model, we added an additional mean-pooling layer and dropout layer on top of the model prior to the final three binary classification layers, each of which corresponding to a class (i.e., 0 (Left), 1 (Center), or 2 (Right)). The highest class probability determines the article’s final classification. This approach is sometimes referred to as the one-against-all approach for multi-class problem [19]. Adding these additional layers has been shown to help prevent over-fitting while fine-tuning. We used an Adam optimizer with a learning rate of $2e - 5$ and an epsilon of $1.5e - 8$. We use a binary cross-entropy loss function, 4 epochs, and a batch size of 32.

Table 5

Accenture results from 2023 CheckThat! Lab Task 3

Task	Classifier Type	Class	Precision	Recall	F1-score
News Article Bias	One-Against-All	0 (Left)	0.822	0.480	0.606
		1 (Center)	0.622	0.791	0.696
		2 (Right)	0.418	0.769	0.542
		macro avg	0.621	0.680	0.615
		weighted avg	0.696	0.633	0.632
News Article Bias	Multi-class	0 (Left)	0.865	0.409	0.555
		1 (Center)	0.599	0.841	0.700
		2 (Right)	0.416	0.783	0.543
		macro avg	0.627	0.678	0.599
		weighted avg	0.709	0.619	0.608
News Media Source Bias	One-Against-All	0 (Left)	0.600	0.720	0.655
		1 (Center)	0.645	0.690	0.667
		2 (Right)	0.634	0.542	0.584
		macro avg	0.626	0.650	0.635
		weighted avg	0.629	0.627	0.625
News Media Source Bias	Multi-class	0 (Left)	0.710	0.880	0.786
		1 (Center)	0.585	0.828	0.686
		2 (Right)	0.867	0.542	0.667
		macro avg	0.721	0.750	0.713
		weighted avg	0.748	0.706	0.701

Table 6

Accenture results from 2023 CheckThat! Lab Task 3

Task	Classifier Type	Accuracy	MAE
News Article Bias	One-Against-All	0.633	0.473
News Article Bias	Multi-class	0.619	0.491
News Media Source Bias	One-Against-All	0.627	0.549
News Media Source Bias	Multi-class	0.706	0.373

For the News Source Bias Classification, the same model architecture above is used to fine-tune a RoBERTa model and classify all article from each news source. We used the majority class label of all articles of a given news source as the class label to establish source bias.

For comparison, we have fine-tuned an additional RoBERTa model for each task above with a single multi-class classifier instead of three binary classifiers. We kept all parameters above the same with a few exceptions: we used the Softmax activation and the sparse categorical cross-entropy loss function instead.

5. Results

Table 5 contains model performance on the test set provided by the organizers. Our One-Against-All News Article Bias had an accuracy of 0.633 and a weighted average F1-score of 0.632. Our One-Against-All News Media Source Bias classifier had an accuracy of 0.627 and a weighted average F1-score of 0.625. The official evaluation numbers are shown in Table 6, where the One-Against-All News Article Bias received an MAE of 0.473 and the One-Against-All News Media Source Bias classifier received an MAE of 0.549.

6. Conclusion

This paper demonstrates models for the classification of political bias in news articles and news sources. In both models, we utilize a pre-trained RoBERTa Large model fine-tuned to the task. We utilized back translation as a strategy to augment the training data, addressing label bias which is found in many machine learning problems. Of the four teams that participated in this shared task, Team Accenture achieved the best overall MAE for 3A (0.473). Of the two teams that submitted, Team Accenture achieved the best overall MAE for 3B (0.549). In addition, for classification of Bias News Article, the one-against-all and the multi-class classifier did not differ significantly in performance. However, for classification of Bias News Sources, the multi-class classifier outperform the one-against-all approach in accuracy but under-perform in MAE. Overall, Team Accenture’s choice of implementing three binary classification layers instead of a single multi-class classifier is the optimal one in both performance and fine-tuning time.

References

- [1] R. R. R. Gangula, S. R. Duggenpudi, R. Mamidi, Detecting political bias in news articles using headline attention, in: Proceedings of the 2019 ACL workshop BlackboxNLP: analyzing and interpreting neural networks for NLP, 2019, pp. 77–84.
- [2] J. Hong, Y. Cho, J. Jung, J. Han, J. Thorne, Disentangling structure and style: Political bias detection in news by inducing document hierarchy, arXiv preprint arXiv:2304.02247 (2023).
- [3] A. Gollwitzer, C. Martel, W. J. Brady, P. Pärnamets, I. G. Freedman, E. D. Knowles, J. J. Van Bavel, Partisan differences in physical distancing are linked to health outcomes during the covid-19 pandemic, *Nature human behaviour* 4 (2020) 1186–1197.
- [4] M. Motta, D. Stecula, The effects of partisan media in the face of global pandemic: How news shaped COVID-19 vaccine hesitancy, *Political Communication* (2023) 1–22.
- [5] R. Baly, G. Karadzhov, A. Saleh, J. Glass, P. Nakov, Multi-task ordinal regression for jointly predicting the trustworthiness and the leading political ideology of news media, arXiv preprint arXiv:1904.00542 (2019).
- [6] W.-F. Chen, K. Al-Khatib, H. Wachsmuth, B. Stein, Analyzing political bias and unfairness in news articles at different levels of granularity, arXiv preprint arXiv:2010.10652 (2020).
- [7] V. Patricia Aires, F. G. Nakamura, E. F. Nakamura, A link-based approach to detect

- media bias in news websites, in: Companion Proceedings of The 2019 World Wide Web Conference, 2019, pp. 742–745.
- [8] P. Hajare, S. Kamal, S. Krishnan, A. Bagavathi, A machine learning pipeline to examine political bias with congressional speeches, in: 2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA), IEEE, 2021, pp. 239–243.
 - [9] E. Kim, Y. Lelkes, J. McCrain, Measuring dynamic media bias, Proceedings of the National Academy of Sciences 119 (2022) e2202197119.
 - [10] Y. Dinkov, A. Ali, I. Koychev, P. Nakov, Predicting the leading political ideology of youtube channels using acoustic, textual, and metadata information, arXiv preprint arXiv:1910.08948 (2019).
 - [11] G. Da San Martino, F. Alam, M. Hasanain, R. N. Nandi, D. Azizov, P. Nakov, Overview of the CLEF-2023 CheckThat! lab task 3 on political bias of news articles and news media, in: Working Notes of CLEF 2023–Conference and Labs of the Evaluation Forum, CLEF '2023, Thessaloniki, Greece, 2023.
 - [12] S. Castelo, T. Almeida, A. Elghafari, A. Santos, K. Pham, E. Nakamura, J. Freire, A topic-agnostic approach for identifying fake news pages, in: Companion proceedings of the 2019 World Wide Web conference, 2019, pp. 975–980.
 - [13] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
 - [14] I. Turc, M.-W. Chang, K. Lee, K. Toutanova, Well-read students learn better: On the importance of pre-training compact models, 2019. arXiv:1908.08962.
 - [15] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A robustly optimized BERT pretraining approach, CoRR abs/1907.11692 (2019). URL: <http://arxiv.org/abs/1907.11692>. arXiv:1907.11692.
 - [16] E. M. Williams, P. Rodrigues, V. Novak, Accenture at CheckThat! 2020: If you say so: Post-hoc fact-checking of claims using transformer-based models, in: L. Cappellato, C. Eickhoff, N. Ferro, A. Névéol (Eds.), Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020, volume 2696 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020. URL: http://ceur-ws.org/Vol-2696/paper_226.pdf.
 - [17] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A robustly optimized BERT pretraining approach, CoRR abs/1907.11692 (2019). URL: <http://arxiv.org/abs/1907.11692>. arXiv:1907.11692.
 - [18] E. Williams, P. Rodrigues, S. Tran, Accenture at CheckThat! 2021: Interesting claim identification and ranking with contextually sensitive lexical training data augmentation, 2021. arXiv:2107.05684.
 - [19] Y. Liu, Y. F. Zheng, One-against-all multi-class svm classification using reliability measures, in: Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005., volume 2, IEEE, 2005, pp. 849–854.