Detecting Sentence-Level Media Bias Using DeBERTa and BABE

John Stilb
UC Berkeley
jm.stilb@berkeley.edu

Abstract

News outlets and their coverage have a strong impact on the public's perception of certain events and topics. However, news outlets and their content can be biased. Automatic detection of such bias faces many challenges, which have been tackled by many different groups. Recently, a team has developed the BABE dataset to develop a "gold standard" for sentence-level bias detection (Spinde et al., 2021 BABE). Using this dataset and a language model pre-training technique known as distant supervision has helped overcome many of the hurdles of effective sentence-level bias detection.

To build upon these results, I have sought to employ a newly developed architecture known as DeBERTa (decoding-enhanced BERT.) This model developed by Microsoft uses disentangled attention and an enhanced mask decoder to tackle natural language understanding and generation tasks more effectively than existing pre-trained models (He et al., 2020). While I was unable to replicate the results of the initial study, I did see a similar performance distribution among the models, if only lower. Additionally, DeBERTa w/ distant supervision outperformed both BERT and RoBERTa on the SG1 dataset, but saw a much lower weighted F1 score on SG2.

1 Introduction

Americans are spending more time online than ever and increasingly consuming news online. Additionally, political polarization has risen dramatically in recent years. Studies have shown that news outlets can be biased in favor of a given party and that this bias can have strong effects on public perceptions.

As people spend more time consuming and sharing increasingly polarized content it has become critical to provide people with tools to better understand the biases of said information. Better bias detection may not only enable smarter information consumption and distribution at the consumer level, but may also empower creators to deploy more neutral language.

2 Background

Bias detection is not without its challenges. Many initial solutions tried to classify entire articles or only headlines (Reddy et al., 2019), but these approaches force articles into a category neglecting that components of the article may differ in terms of bias. More recent approaches focus on sentence-level bias classification.

Another major issue is that bias is difficult to define and agree upon. Crowd-sourced datasets such as MBIC (Spinde et al., 2021 MBIC) have sought to tackle this problem by allowing for many people to decide whether there is bias in a given example. However, the variance amongst conclusions is high. In an effort to reduce this variance, a team created the BABE dataset.

BABE (bias annotations by experts) is a dataset of media bias annotations generated by experts using a more defined methodology. It was generated by extracting sentences from news articles that were focused on 12 controversial topics (Spinde et al., 2021 BABE). The resulting dataset shows less variance amongst the labels than that of MBIC. Even with the improved dataset, BABE still has one major shortcoming, which is its size.

The team that generated BABE hopes to overcome the small size of BABE by using a technique known as distant supervision. This pre-training technique uses "distant" (noisier) data to incorporate information from a given task into the initial

weights of a downstream model. The primary benefit being, it allows the model to improve feature extraction on a given task by being introduced to more—albeit noisier—relevant data.

Using the BABE dataset and the distant supervision technique, I have sought to improve bias detection model performance using a new language model architecture known as DeBERTa. DeBERTa (decoding-enhanced BERT) improves upon BERT and RoBERTA by using a disentangled attention mechanism and an enhanced mask decoder (He et al., 2020). These new techniques improve model pre-training and performance in natural language understanding and generation tasks. Given these improvements, I want to apply the same architecture to the BABE (bias annotations by experts).

3 Methods

I propose to mimic the process used in (Spinde et al., 2021 BABE) by using a variety of language models to solve the media bias classification task. Their methods include using a distant supervision framework to pre-train the feature extraction algorithms to improve the language models representation of the data by including information about a sample's bias. Additionally, they chose to use noisy labels as it is more abundantly available than human-labeled data while still providing supervisory signals.

3.1 Learning Task

Given a randomly sampled sequence of tokens in a corpus, my task is to assign the correct label to a given sequence where 0 represents the neutral class and 1 represents the biased class. I optimize this task through the minimization of the binary cross entropy loss function.

3.2 Dataset Evaluation

Label	SG1	SG2
Biased	43.88%	49.26%
Non-Biased	47.01%	50.71%
No Agreement	9.01%	0.03%

Table 1: Bias Label Distribution.

Our primary models are trained on two different datasets: SG1 and SG2. SG1 consists of 1,700 sentences annotated by eight experts each. SG2 consists of 3,700 sentences annotated by 5 experts each. Both datasets use a majority vote principle to decide whether a sentence is biased.

In Table 1 we can see that SG2 saw much more agreement and annotators were less hesitant to label a sentence biased. Additionally, SG2 saw many more sentences as factual than opinionated versus SG1, but both datasets have more factual sentences than opinionated sentences as seen in Table 2. Given the larger number of annotators working on SG1, it makes sense that we'd see less agreement.

Label	SG1	SG2
Opinionated	25.00%	23.35%
Factual	37.59%	43.55%
Both	26.65%	27.22%
No Agreement	10.76%	5.88%

Table 2: Opinion Label Distribution.

Additional information about the distribution of topics, bias per topics, bias/ideology distribution can be found in the eda notebook in the provided repository.

3.3 Language Models

I replicated the language models used in the (Spinde et al., 2021 BABE) paper to see if I'm able to replicate the results in the paper as well as improve upon them through the use of the DeBERTa model. They used a variety of models including BERT and one of its variants RoBERTa. These models used unlabeled text to create bidirectional representations of language. The team also uses ELECTRA which learns language representations through discrimination.

DeBERTa, another BERT variant, differs from the other models through its use of a disentangled attention mechanism as well as its use of an enhanced masked decoder (He et al., 2020). In DeBERTA each word is represented using two vectors that encode its content and relative position. The attention weights among words are computed from these vectors using disentangled matrices.

The masked decoder incorporates absolute word position embeddings before the softmax layer so absolute position is taken into account as well as relative position that is incorporated in the disentangled matrices. Additionally, it uses a virtual adversarial training method for fine tuning that improves generalization of the language model for downstream tasks.

3.4 Distant Supervision

(Spinde et al., 2021 BABE) introduced additional pre-training before fine tuning to improve feature learning with regards to media bias content. In this stage, they incorporate bias information directly in the loss function to remedy the lack of information on language bias in the embedded space of the pre-trained models.

This stage consists of predicting "distant" or "weak" labels from a larger, more noisy dataset. This pre-training dataset consists of news headlines from media outlets with and without partisan leaning. All headlines from a given outlet will have the same "bias". To guarantee model integrity they ensure there is no overlap between the pre-training data and the BABE dataset.

3.5 Implementation

All language models are instantiated with their pre-trained parameters. Batch sizes are adjusted to overcome memory limitations, but the buffer size is consistent across all models. Each model uses the Adam optimization with a learning rate of 5e-5. For the baseline, I ran each model for only a single epoch, but trained the top performing models on up to 10 epochs. I use a mechanism to stop training when there is no improvement to loss and restore the best weights.

3.6 Evaluation

Since BABE is a small dataset consisting of only 3,700 sequences and since SG1 is unbalanced, I report performance metrics using 5 fold stratified cross validation to stabilize results and maintain this imbalance in each fold. I report final scores using a weighted average of F1-scores.

4 Results and Discussion

4.1 Additional Pre-Training Performance

Model	Loss	Accuracy
BERT	0.4057	0.7711
RoBERTa	0.2819	0.6545
DeBERTa	0.2751	0.6457

Table 3: Distant Pre-training Results.

The goal of the distant supervision additional pre-training step was to better initialize our models by first training them on noisier, but larger datasets related to our final classification task. The additional pre-training of the models saw relatively strong results across the board as seen in Table 3. DeBERTa saw the lowest accuracy scores, but minimized loss more than the other 2 models. BERT on the other hand, saw the highest accuracy score, but also the highest loss.

4.2 Model Performance

Model	SG1	SG2
BERT	**0.413136**	0.378480
RoBERTa	0.377869	0.419941
DeBERTa	0.389934	0.335278
BERT w/ Distant	0.320406	**0.502989**
RoBERTa w/ Distant	0.378031	**0.522579**
DeBERTa w/ Distant	0.396586	0.310578
ELECTRA	**0.448213**	0.440985

Table 4: Stratified 5 fold cross-validation Weighted Macro F1 Scores.

My results, as seen in Table 4 were lower across the board from the original paper (Spinde et al., 2021 BABE), despite my close adherence to their methodology. For the 1st dataset (SG1,) ELECTRA performed the best, followed by BERT. Interestingly, of the models which had additional pre-training BERT was the only one unable to outperform its base model. DeBERTA w/ distant supervision was the third highest performing model on SG1, but performed the worst on the second dataset (SG2.)

On the second dataset, the top two performing models used distant supervision, with RoBERTa performing the best, followed by BERT. Despite RoBERTa being an essentially larger version of BERT, it performed worse on the first dataset, but better on the second. In the initial paper, all models generally performed better on the second dataset due to its larger corpus size. Our difference likely stems from variability in training due to the small size of the datasets.

4.3 Analysis and Discussion

To explore why the models performed as they did, I ran several sentences through several of our key models. I chose a random sentence for each opinion label and bias label pairing for both datasets to better understand model performance on different sentence types.

Out of the sampled SG1 sentences, BERT only made predictions of "Bias" for those with an opinion label of "Expresses writer's opinion". BERT w/ distant supervision accurately predicted every value, regardless of label, which is in contrast to its overall performance indicating a favorable sample. DeBERTa heavily favored making no prediction in SG1, while RoBERTa and ELECTRA were fairly balanced.

Out of the sampled SG2 sentences DeBERTa w/ distant supervision and BERT w/ distant supervision tended to make the bias prediction, while the others favored the "non-biased" prediction. There does not appear to be strong correlations between the opinion label and our predicted value in SG2, which aligns with the increased agreement by the annotators.

Model	% Bias	% Non-Bias
BERT	33.33	66.67
DeBERTa	8.33	91.67
BERT w/ Distant	58.33	41.67
RoBERTa w/ Distant	25.00	75.00
DeBERTa w/ Distant	50.00	50.00
ELECTRA	41.67	58.33

Table 5: Prediction Distribution Across Sampled Sentences.

In Table 5, we can see that overall DeBERTa generally prefers to make no "bias" prediction, which explains its better performance in SG1 which is far less balanced than SG2 as seen in Table 1. However, adding distant supervision improved the balance of its predictions. In fact, across every model that had distant supervision, I saw a better balance of predictions. Given the small size of our datasets, it makes sense that we'd see more robust predictions with additional pre-training. ELECTRA saw fairly stable prediction performance, which reflects its strong performance across both datasets.

Initially, I expected DeBERTa to outperform both BERT and RoBERTa due to its ability to account for relative and absolute positions of words in a given sentence. It seems when making predictions of bias, that this positional information does not provide as much value and may in fact detract from overall performance.

5 Conclusion

I was unable to replicate the results of (Spinde et al., 2021 BABE), but I did see a similar distribution of performance. These differences likely represent variance stemming from the small size of the dataset and selection of hyper-parameters during optimization. While our results do differ, I do conclude that SG1 and SG2 are superior datasets compared to their predecessors such as MBIC.

Given its performance and my analysis of its performance, DeBERTa appears to be a weak fit for bias prediction, likely because the inclusion of absolute and relative positional information is not as important in the detection of bias at the sentence level.

References

Rama Rohit Reddy, Suma Reddy Duggenpudi, Radhika Mamidi 2019. *Detecting Political Bias in News Articles Using Headline Attention*. International Institute of Information Technology, Hyderabad Proceedings of the Second BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, pages 77–84 Association for Computational Linguistics

He, Pengcheng and Liu, Xiaodong and Gao, Jianfeng and Chen, Weizhu 2020. *DeBERTa: Decoding-enhanced BERT with Disentangled Attention*. Microsoft Dynamics 365 and Microsoft Research arXiv

T. Spinde, L. Rudnitckaia, K. Sinha, F. Hamborg, B. Gipp, K. Donnay " 2021. *MBIC – A Media Bias Annotation Dataset Including Annotator Characteristics*. In: Proceedings of the iConference 2021.

Timo Spinde, Manuel Plank, Jan-David Krieger, Terry Ruas, Bela Gipp, and Akiko Aizawa. 2021. *Neural Media Bias Detection Using Distant Supervision With BABE - Bias Annotations By Experts*. University of Wuppertal, University of Konstanz, NII Tokyo In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 1166–1177, Punta Cana, Dominican Republic