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. This model developed by Microsoft uses disentangled attention and an enhance mask decoder to tackle natural language understanding and generation tasks more effectively than existing pre-trained models (He et al., 2020). Thus far, I have failed to replicate the results of the BABE paper and DeBERTa has actually performed worse than existing models on the task of sentence-level bias detection.

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 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 headlined (Reddy et al., 2019), but this approach forces 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 super-

vision technique, I have sought to improve bias detection model performance using a knew language model architecture known as DeBERTa. DeBERTa improves upon BERT and RoBERTA 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 Language Models

I replicate 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 some of its variants of DistilBERT and RoBERTa. These models used unlabeled text to create bidirectional representations of language. Additionally they use ELECTRA which learns language representations through discrimination and XLNet which is an auto-regressive model.

DeBERTa, another BERT variant, differs from the other models through its use of a disentangled attention mechanism as well as its 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 disentangle 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.

3.3 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.

3.4 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 only ran each model for a single epoch, but will train 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.

3.5 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.

Model	Loss	Accuracy
BERT	0.4790	0.7403
RoBERTa	5.4643	0.6457
DeBERTa	0.6505	0.3543

Table 1: Distant Pre-training Results.

4 Results and Discussion

4.1 Distant Pre-training

The pre-training of the models saw relatively strong results for BERT and RoBERTa, but DeBERTA struggled to improve accuracy. Despite this all models using pre-training saw worse performance than their respective model without pre-training with the exclusion of "DeBERTa w/ distant" on the SG2 dataset. However this model saw an F1 score of 0.

4.2 Model Performance

Model	SG1	SG2
BERT	0.420495	0.414778
RoBERTa	0.377873	0.401316
DeBERTa	0.337098	0.334656
BERT w/ Distant	0.373501	0.324109
RoBERTa w/ Distant	0.328904	0.338353
DeBERTa w/ Distant	0.32993	0.34138
DistilBERT	0.440871	0.445587
ELECTRA	0.441015	0.463769
XLNet	0.367079	0.320916

Table 2: Stratified 5 fold cross-validation Macro F1 Scores.

My results varied widely from those in the original paper (Spinde et al., 2021 BABE). In my initial experiments, DistilBERT and ELECTRA outperformed BERT and RoBERTa. Additionally, my results are much lower than those of the original paper across the board. This likely stems from the smaller batch size and number of epochs run.

5 Next Steps

Due to poor performance of the distant-supervised models, I plan to rework the pre-training of these models. Additionally, I will continue to fine tune the top performing models along with DeBERTa to improve overall performance. I will investigate why DeBERTa has under performed so we can better understand the implications of the model's unique attributes. Lastly, I will run the top performing models on both BABE and MBIC to generate our final results.

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