Political Ideology Bias Detection with BERT

Ahsen Qamar and Alex Dauenhauer

August 3, 2019

Abstract

In this paper, we attempt to use Google's recently released BERT model to detect political ideology bias at the sentence level. We use a pre-processed dataset from the Ideological Books Corpus (IBC) [3] as well as adapting a method of filtering and labeling from Iyyer et al. [3] to generate two new datasets from raw document corpora, the Convote dataset [5] and the All-the-news dataset [6]. We perform experiments using a "train on one, test on all" format. Our model performs well when evaluated on the same set it was trained with, and shows that there is room for improvement in being able to generalize from one dataset to the other. Our model achieves F1-scores of 99.4% when trained and tested with our All-the-news dataset, 92.6% with the Convote dataset and 56.8% with the IBC dataset. For document level bias detection, we achieve an F1-score of 60% on the Convote dataset and 50% on the All-the-news dataset.

1 Introduction

Political ideology bias in news sources is a topic of growing concern, not just in the U.S., but across the entire world. As our country grows ever more partisan and misinformation campaigns fuel distrust in mainstream news sources, people have a tendency to turn to alternative news sources which typically reflect the existing ideology bias of the author. This creates echo chambers and reinforces existing partisan ideologies driving the partisan divide ever wider. The ability to automatically label biased text could inform readers that may have otherwise assumed the content was neutral, and encourage them to pursue alternative sources with less bias, or at the very least maintain a degree of skepticism about the content they are consuming. A system like this could also help media sources reduce the amount of bias they include their reporting, some of which may be unconscious and therefore undetected by writers or editors. Using this tool as a proofreader of sorts could eventually, hopefully reduce the amount of bias in media and reduce the growing partisan divide.

In this paper, we build off of previous work done by Iyyer et al. [3]. Their success in using an RNN model to predict ideological bias at the sentence level, inspired us to expand on their experiments using BERT [1], a modern transformer-based model that achieves state-of-the-art results[1] on a number of different NLP problems. We use a BERT implementation for sequence classification to predict ideological bias at the sentence level on three unique datasets. We use a processed subset of sentences from the Ideological Books Corpus (IBC) [2]. We also use a filtering method explained in section 3.4 to select a set of biased sentences from the Convote dataset [5]. Additionally, we build a new dataset by filtering a corpus of news articles from 15 different publishers [6], with bias labels sourced from mediabiasfactcheck.com (MBFC) [8]. We predict bias at the sentence level, then use the aggregate sentence level predictions to predict bias at a document level and compare to the publisher label assigned by the MBFC team. These datasets are described in greater detail in section 3.

2 Model & Architecture

BERT (Bidirectional Encoder Representations from Transformers) was released late last year by Google. The main draw towards the use of BERT for this project was leveraging the bi-directional training transformer to help detect political bias in text. For more information regarding BERT please refer to this research paper from Google.

A cloud instance hosting an Nvidia P100 GPU was our environment for hosting our data as well as deploying our model for training and evaluation. This infrastructure limited us to use the BERT *base* models as the GPU would continually run out of memory with the slightly larger models. No significant difference was found between *cased* and *uncased* for the *base* models so we decided to move forward with *BERT-Base, Cased*. The decision

[1] At the time this sentence was written

was made to move forward with a **pytorch** backend due to more thorough availability of code examples and better documentation. The primary reference for our code was from a *Medium* article [4]. A high level diagram of our overall architecture is shown in Figure 1. Data was obtained from three different sources and funneled

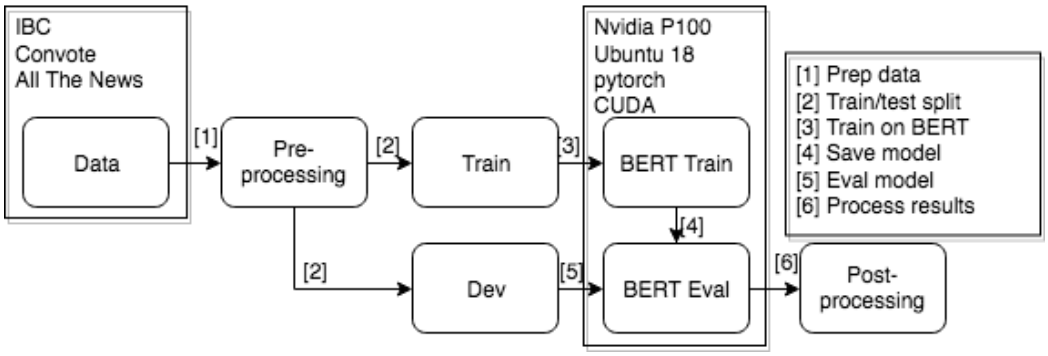

Figure 1: Architecture

into a pre-processing step. This involved generating sentences with labels of liberal, neutral, or conservative and formatting the data into a tsv that BERT would accept. We then perform a random 70:30 train/dev split. Train data is fed into our BERT train script and the generated model is saved and funneled into our BERT eval script. The model is then evaluated on the dev data. An averaging classification layer is added as part of post-processing for documents. Let's take a look at what all is moving through our pipeline.

3 Data

Political bias at the sentence level is a very subjective topic and therefore, there are not many large corpora widely available for use. A large part of this project was devoted to acquiring and preparing a few existing corpora, as well as repurposing a few processing methods on new corpora to develop our own biased sentence corpus. We performed experiments on three separate datasets: the Ideological Books Corpus (IBC) [3], Annotated congressional floor debates [5], and the all-the-news dataset from Kaggle user Andrew Thompson [6]. In this section we will describe the content of each dataset and processing steps performed on each dataset for use in our model.

3.1 IBC

The Ideological Books Corpus was provided to us in a fully processed format, courtesy of Iyyer et al.[3] The original IBC dataset, developed by Gross et al.[2] is a collection of books and articles written between 2008 and 2012 by well-known authors with strong political leanings. What Iyyer et al. did was to filter this corpus using a similar strategy to our strategy outlined in section 3.4[2]. They then crowd-sourced manual ideological bias annotations of the resulting sentences and particular subphrases. We use the data in this processed form as is, with no further processing. This provides a more subjective approach to bias labeling which we can compare with our objective labeling approach described in section 3.4.

3.2 Convote

The Convote dataset is a corpus of congressional speeches with each speech treated as a document with automatically derived labels of the speaker's political party (D, for Democrat; R, for Republican; I for Independent) as well as other related extracted information that was not pertinent to our experiments. Since our work attempts to predict ideological bias rather than political party, we relabel each document by mapping Democrat to "liberal", Republican to "conservative" and Independent to "neutral"[3]. While it is true that the mapping of political party directly to political bias is not always a 1:1 relationship, there is a strong correlation between political party and political ideology. Further, we expect that our method for filtering the dataset (explained in section 3.4) for

[2]Iyyer's team used bigrams and unigrams as their bias detectors, whereas we use trigrams and bigrams

[3]While we realize that a politician of Independent party certainly does not imply that politician's dialogue will contain no ideological bias, we use this label to select sentences that contain no bias detectors from either the liberal or conservative ideology which is explained further in section 3.4.

biased sentences will wash out noise that would be seen from speeches made by moderate centrists on either side of the aisle.

3.3 All-the-news Kaggle corpus

To expand our training data further with a greater diversity of authors, we turned to a Kaggle corpus of 142,570 news articles from 15 different publishers as provided by Andrew Thompson [6]. We assigned each publisher a bias label, sourced from mediabiasfactcheck.com (MBFC), then we simplified these labels down to the same labels we used in the previous datasets: liberal, conservative and neutral. The labels assigned to each publisher are shown in Table 1.

Table 1: Publisher and bias labels from all-the-news corpus

Publisher	MBFC Bias Label	Simplified Bias Label
New York Times	left-center	liberal
Breitbart	extreme-right	conservative
CNN	left	liberal
Business Insider	left-center	liberal
Atlantic	left-center	liberal
Fox News	right	conservative
Talking Points Memo	left	liberal
Buzzfeed News	left-center	liberal
National Review	right	conservative
New York Post	right-center	conservative
Guardian	left-center	liberal
NPR	left-center	liberal
Reuters	neutral	neutral
Vox	left	liberal
Washington Post	left-center	liberal

3.4 Biased Sentence Selection

It would be extremely unreasonable to assume that every sentence spoken by a member of congress during a congressional debate would contain ideological bias. Or that every sentence written by every author published under a specific news publisher would contain the bias of that publisher. In fact, a large portion of the sentences in both the Convote dataset and the All-the-news dataset contain no bias at all. Therefore, it is necessary to filter these datasets for sentences that explicitly contain bias, prior to labeling those sentences and using them to train our model.

To select the explicitly biased sentences from the data, we used a method inspired by Yano, et al. [7] and similar to a method used by by Iyyer, et al [3] that was shown to be successful at identifying bias indicators. We started by identifying the most frequently used trigrams and bigrams for each bias label, liberal or conservative (in the Convote set, we used the full corpus to identify n-grams, whereas in the All-the-news data, we use a subset of the publishers with the most extreme MBFC bias labels to identify n-grams. This subset of publishers is highlighted in gray in Table 1). We then filtered out any trigrams or bigrams which contained stopwords, and English names[4]. We further filtered each bigram list by requiring that at least one word in each bigram contain an "opinion" word, defined as a word found in the `opinion_lexicon` corpus from NLTK[5]. We then took the set difference of the resulting top 1000 most frequent liberal and conservative n-grams. We kept the top 100 filtered bigrams and top 100 filtered trigrams for each label as our bias indicators (for a total of 200 bias indicators for each label). We then filtered sentences such that, if a democratic speaker (or author from a liberal publisher) spoke a sentence that contained one of these top 200 liberal bias indicators, we labeled that sentence as "liberal"

[4]Removing stopwords serves the purpose of eliminating trigrams and bigrams that contain commonly used words, but don't contain substance for bias. Removing English names was a strategy that proved effective in filtering the All-the-news dataset (removing references to prolific journalists or prominent politicians), and was applied to the Convote data as well, essentially serving the purpose of removing n-grams that are sourced from replies to other speakers that address the other speaker by name

[5]This strategy intends to select only those phrases that contain an ideological opinion which would typically be seen as a strong indicator of bias, when spoken by someone with known political affiliation. We decided to not make this a requirement for the trigrams as we felt these phrases would be unique enough to each ideology without this additional criteria that we did not want to lose information by filtering too heavily.

and used the same logic for republican speakers/conservative publishers to identify "conservative" sentences. For these datasets, we identified a neutral sentence as one which was spoken by an independent politician or neutral publisher[6] and contained no n-grams from either the liberal or conservative bias indicator lists. The top 10 bias indicators from each label for the Convote data are shown in Table 2 and those from the All-the-news data are shown in Table 4. The resulting Convote dataset included 1326 conservative sentences, 1383 liberal sentences and 213 neutral sentences. The resulting All-the-news dataset contained 16,745 conservative sentences, 28,230 liberal sentences, 277,751 neutral sentences. To balance this dataset and avoid weighting strategies while training, we pared this down to a random selection of 16,000 sentences from each label, for a total of 48,000 sentences. Example sentences from each ideology label for the Convote data are shown in Table 3 and examples from the All-the-news data are shown in Table 5.

Table 2: Top 10 n-grams per ideology label - Convote

Liberal		**Conservative**	
Trigrams	Bigrams	Trigrams	Bigrams
social security trust	tax breaks	national electrical contractors	community protection
security trust fund	security trust	electrical contractors association	free market
cbc alternative budget	bad policy	legislative days within	organized crime
black caucus budget	would lose	inner cell mass	bankruptcy relief
estate tax relief	reduce crime	head start program	good news
privatize social security	budget reconciliation	community protection act	relief extension
u.s. trade deficit	ethical standard	million new jobs	delayed notification
republican budget resolution	fiscally irresponsible	death tax repeal	soft money
national wildlife refuge	working poor	9/11 commission report	illegal aliens
guardian ad litem	subpoena power	stem cells without	invasive species

Table 3: Sample sentences from each ideology label - Convote

Liberal	**Conservative**	**Neutral**
mr. speaker, during a time of war, in the aftermath of a catastrophic hurricane, with 45 million americans lacking health insurance and skyrocketing home heating costs projected this winter, this majority is proposing to take from those with the least, give to those with the most – and tell our children they will have to pay for it all later.	on both the business records and delayed notification sections of the patriot act (among others), the stance of the american civil liberties union and like-minded critics seems to have an ulterior motive.	let us look at what is going on in america today.
it is clear that there would be plenty of money to deal with the social security trust fund if the president were not using the social security trust fund as a slush fund to give tax cuts to the wealthiest people in america.	that legislation helped to streamline the intelligence community and tightened some asylum rules that allowed potential terrorists to remain in our country.	mr. speaker, parliamentary inquiry.

[6]Reuters is the only publisher labeled as "neutral" by MBFC

[7]The extreme offensiveness of this bigram led us to investigate its origins further. The presence of this bigram is not actually due to conservative media using slurs like this commonly enough to rank in our top ten list. Instead it is actually a reference to "The Dangerous Faggot Tour" which is a campus speaking tour of Milo Yiannopoulos, a prominent contributor to Breitbart News. This is not to condone the use of this offensive slur, just to explain its origins and presence in this list a bit more clearly

Table 4: Top 10 n-grams per ideology label - All-the-news

Liberal		**Conservative**	
Trigrams	Bigrams	Trigrams	Bigrams
senior administration official	opioid epidemic	jerusalem bureau chief	illegal aliens
greenhouse gas emissions	health reform	border patrol agent	illegal alien
north korean leader	lethal injection	social justice warriors	illegal immigrant
federal civil rights	healthy people	black panther party	migrant crisis
provocative narrative essays	racial bias	cartel chronicles project	patriot channel
clean air act	budget reconciliation	refugee resettlement program	hard truths
health care policy	chronic pain	prison sentence commuted	twin falls
republican health care	lead poisoning	god less america	islamic terror
gop health care	intelligence committees	real clear politics	snarky opinions
civil rights laws	rights advocates	face certain death	dangerous faggot[7]

Table 5: Sample sentences from each ideology label - All-the-news

Liberal	**Conservative**	**Neutral**
Here's what you need to know: American divisions are rapidly widening over President Trump's order to close the U. S. to refugees and people from seven predominantly Muslim countries.	And the costs of illegal alien crime continued to mount and a lethal opioid epidemic raged.	Showcasing their attempts to unite with other groups for the election, Islamists campaigned with Awdeh Qawwas, a prominent priest, in the affluent Abdoun district of the capital Amman.
In a video posted on her campaign's Facebook page shortly after Mr. Sanders departed the White House grounds to visit the Capitol, Mr. Obama described Mrs. Clinton as the most qualified candidate to seek the White House, and implored Democrats to come together to elect her after a divisive party primary.	Obama's claim of civic peace is also at odds with the televised evidence: dramatic race riots, cop killings, rapes, murders, illegal alien crimes, and chaos that rippled across the country during the second term of his presidency.	Rousseff's survival hinges on winning over a dwindling number of undecided lawmakers who are also being courted by the man poised to take over if she is ousted, Vice President Michel Temer.

4 Experiments/Results

When experimenting with BERT we found the following configuration of BERT parameters to provide the best results: $MAX\_SEQ\_LEN = 256$, $TRAIN\_BATCH\_SIZE = 16$, $LEARNING\_RATE = 2e-6$, $NUM\_TRAIN\_EPOCHS = 3$. The MAX_SEQ_LEN and $TRAIN\_BATCH\_SIZE$ are the maximum possible values selected given we were using an Nvidia P100 and concurrently dependent on our Cuda version. The $LEARNING\_RATE$ is the default provided by BERT. NUM_TRAIN_EPOCHS was set to 3 after continuous experimentation and analysis of the evaluation loss which saturated after about 2.5 epochs. These parameters remained the same for all three datasets.

As seen in Table 6, the *IBC* dataset had the worst performance both on itself as well as the other datasets. The evaluation accuracy is highest on itself, followed by *AllTheNews* with an accuracy of 40%, and then *Convote* with an accuracy of 40%. What is surprising is that f1 scores are higher than accuracy on all the datasets besides itself. The evaluation loss saturated at 0.99 (Table 7) hinting at the fact that the dataset is fairly small.

The *Convote* had fairly good performance with an evaluation accuracy of 91% on itself. It had an accuracy of 63.5% on *AllTheNews* which is significantly better than *IBC*. Lowest evaluation accuracy was on *IBC* at 20%. F1 scores very closely mirrored their respective accuracies. The evaluation loss improved when compared to *IBC*

at 0.23. Again, the evaluation loss saturated indicating we still might not have enough data.

The best performance of all the datasets was from *AllTheNews* with an evaluation accuracy of 99.4% on itself. It had a similar evaluation accuracy on *IBC* to that of *Convote*. On the *Convote* dataset it obtained an evaluation accuracy of 60%. The f1 scores were lower on the *IBC* and *Convote* when compared to the evaluation accuracies. We saw the lowest evaluation loss at 0.02 which is significantly better than any of the other datasets. Switching to document-level classification[8] we see both *Convote* and *AllTheNews* dataset have similar accuracies at about 60%. However, the *AllTheNews* dataset has an f1 score that is 15% lower. This could mean that the model trained on this dataset is struggling with precision and/or recall.

Table 6: Results Matrix

Train/Test	**IBC**		**Convote**		**AllTheNews**		**Documents**	
	Accuracy	**F1-score**	**Accuracy**	**F1-score**	**Accuracy**	**F1-score**	**Accuracy**	**F1-score**
IBC	60%	56.8%	22.8%	28.6%	40%	49%	N/A	N/A
Convote	20%	16.8%	91%	92.6%	63.5%	63.4%	61.18%	60%
AllTheNews	16.8%	10.5%	60%	49.1%	99.4%	99.4%	65%	50%

Table 7: Evaluation loss

IBC	**Convote**	**AllTheNews**
0.99	0.23	0.02

5 Conclusion

5.1 Bias labeling at phrase level

The n-gram filtering task shed light on how each ideology uses different, yet equally charged phrases to describe the same term. For example, one of the top liberal trigrams in the Convote dataset was "estate tax relief" whereas a top conservative trigram was "death tax repeal", where both are referring to the same topic. Another example, in the all-the-news dataset, a top conservative bigram is "illegal immigrants" with the liberal equivalent being "undocumented immigrants"[9]. In the Convote and All-the-news datasets, we used the presence or absence of an n-gram bias phrase in a sentence to label the ideology bias at the sentence level. This provides an objective methodology for bias labeling, rather than a subjective system, which would be heavily dependent on the reader's own personal biases. The IBC corpus relies on a subjective (human-determined) labeling methodology[10].

5.2 Bias detection at sentence level

The first and foremost conclusion we can draw from experiments to detect bias at the sentence level is that our model performs well when evaluated using a test set from the corpus with which it was trained (the diagonal of Table 6), but it is not generalizing well enough to handle the variation in language style that exists between our datasets. For example, when we train the model with the all-the-news corpus, we achieve an F1-score of 99.4% when using a test set from the same corpus, but testing on the IBC corpus gives an abysmal F1-score of 10.5%. Training on the IBC corpus (which consists of sentences extracted from books written by prominent authors with known ideological bias) performs terribly on the Convote corpus, and is no better than a coin toss on the all-the-news corpus. This isn't completely surprising since there are great differences between the language used during a parliamentary speech and the language used in an ideological book (which may be trying to persuade its reader to believe a certain idea), which is in turn very different from the language used in a news article (which has limited words to convey its message and in theory should be trying to present the facts of an event, rather than opinions about an event).

[8] Our classification classes were defined as: Liberal – 2, Neutral – 1, Conservative – 0. Our document classification was defined as: Conservative – 0-0.49, Neutral – 0.5-1.49, Liberal – 1.5-2.0

[9] This was one of the top bigrams in content labeled "liberal", however, neither the word *undocumented*, nor the word *immigrants* are an opinion lexicon and therefore our final filtering strategy removes this bigram as a bias detector. In future work, we would like to determine a more sophisticated n-gram detection method

[10] While the crowd-sourced nature of the labeling of this corpus removes some measure of subjectivity, the labels are still based on human opinions of whether a sentence is biased or not.

The drop in performance we see with the IBC dataset as compared to the Convote and All-the-news datasets is likely explained by greater variation in sentence content, relative to the size of the dataset. The IBC dataset is labeled in a subjectively (by humans) rather than the objective labeling method we applied to the Convote and All-the-news data. This allows for a much broader interpretation of what content would consitute bias. In the objective case, the sentence is required to contain one or more n-grams from a curated list to be labeled as biased which restricts the allowable content and subject matter of the sentence. In the subjective case, there is no such requirement and the human labels the bias of the sentence based on their own judgement[11]. We expect that expanding the size of this dataset would improve the performance, as the variation in sentence content should remain relatively constant, but having more samples would allow the model to learn the patterns that define bias more effectively.

5.3 Bias detection at document level

In any given document, a vast majority of the sentences will likely be unbiased, or neutral. Therefore, an interesting question is how well we can detect the bias of the author, given the content of a document. These results are represented in column *Documents* in Table 6. Our results show that our method for document level bias detection (averaging the predicted class for all sentences in the document) has room for improvement. We believe using a Hierarchical Attention Network would drastically improve results at this level, by allowing the sentence level attention vector to have greater and more accurate influence over the resulting document level output.

Future Research

To expand on this project in the future, we would like to experiment with different bias detection methods in the source data. Use of n-grams with $n > 3$, alternative filtering methods such generating an opinion list customized to our dataset, or possibly investigating other techniques from the field of political science.

To further develop this model and improve the usefulness of it at the document level, we would also like to incorporate BERT as an encoder in a Hierarchical Attention Network (HAN). In the most simplistic description, the HAN model encodes words into word vectors, then uses an attention mechanism to aggregate these into a sentence vector. Then the model repeats the process using the sentence vector with an attention mechanism to aggregate the sentences into a document vector for classification. We would like to incorporate BERT as our encoder in the HAN architecture which we believe would improve performance at the document level. We also will need to drastically improve the bias labels at the document level from the datasets we currently have available. This may involve a crowdsourcing task. Unfortunately we were not aware of the HAN model at the start of this project and when we found the architecture we did not have enough time to broaden our scope and incorporate it into this project.

[11] Iyyer et al. also used a similar n-gram filtering method to select sentences that would be labeled by humans, but the filtering method they used was far less restrictive than ours (which is ok since the sentences will be labeled by humans after filtering). Additionally, they smoother the subjectivity of the labeling by crowd sourcing the labels, but this still leaves an element of subjectivity to the labeling of the data

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[2] Justin Gross, Brice Acree, Yanchuan Sim, and Noah A. Smith. Testing the Etch-a-Sketch hypothesis: A computational analysis of Mitt Romney's ideological makeover during the 2012 primary vs. general elections. In *APSA 2013 Annual Meeting Paper*, Chicago, IL, USA, 2013.

[3] Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. Political ideology detection using recursive neural networks. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1113–1122, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[4] Thilina Rajapakse. A simple guide on using bert for binary text classification. https://medium.com/swlh/a-simple-guide-on-using-bert-for-text-classification-bbf041ac8d04, 2019.

[5] Matt Thomas, Bo Pang, and Lillian Lee. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. *CoRR*, abs/cs/0607062, 2006.

[6] Andrew Thompson. All the news: 143,000 articles from 15 american publications. https://www.kaggle.com/snapcrack/all-the-news, 2017.

[7] Tae Yano, Philip Resnik, and Noah A. Smith. Shedding (a thousand points of) light on biased language. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 152–158, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[8] Dave Van Zandt. Media bias/fact check: The most comprehensive media bias resource. https://mediabiasfactcheck.com/, 2015.