BERT for Summarization: Sentiment Analysis for Identifying Gender Bias

Isaac Chau, Alex S Kim

Abstract

Social biases in NLP models have been recognized as a problem for many years. While the rise of excellent pre-trained language models has accelerated research and expanded commercial applications, the consolidation of the field around just a handful of popular models has raised not just the possibility of broadening the reach of social biases that are built into models but also the possibility of mitigating bias in NLP on a wide scale. We applied a simple word-sentiment model to the input/output pairs of a summarization model based on the popular BERT pre-trained language model to search for evidence of gender bias. We found evidence of minor gender bias in the model's word choices. This work is an example of one approach to the growing sub-field of addressing social biases in NLP.

1 Introduction

Building models for NLP tasks that approach, rival, or even set benchmarks for state of the art performance has been made accessible to anybody by the rise of pre-trained language models. Since 2019, the best known and most widely used pre-trained model has been BERT, short for Bidirectional Encoder Representations from Transformers (Devlin et al.2019). While a plethora of research has been published regarding the applications of BERT, there exists far less literature on the well-documented problem of social biases that can be found in NLP models (Bolukbasi et al., 2016), although the recognition of the need to address this issue seems to be growing (Webster et al., 2019). As the popularity of the same handful of pre-trained language models grows, it becomes more important to monitor them for bias. In this paper, we search for evidence of gender bias introduced by the popular BERT pre-trained language model on generative text summarization.

2 Methods

Summarization Model

For our BERT-based text summarizer, we chose the BertExtAbs model from Yang and Lapata (2019) in which they combined a standard encoder-decoder framework for abstractive summarization (See et al., 2017) with a two-stage fine-tuning approach, first fine-tuning the encoder on the extractive summarization task followed by fine-tuning it on the abstractive summarization task. This model, trained on the CNN/Dailymail dataset (Hermann et al., 2015), achieved state-of-the-art results in both automatic and human evaluations. We selected this model for our analysis due to its state-of-the-art performance and its availability as a pre-trained model, as the authors have made it available for download and use. The pre-trained model was necessary for us due to a lack of computing resources sufficient to train our own model on an adequately-sized dataset in the timeframe of our project.

In addition to logistical reasons, we chose the BertExtAbs model due to its combination fine-tuning method. If we had examined the output of a purely extractive text summarizer, our sentiment analysis would have been hindered by the fact that the outputs consist of sentences written by the authors of the original example texts, making it more difficult to separate the sentiment in the text created by humans and that of the model. On the other hand, summaries created by purely abstractive

systems have a greater propensity to produce "disfluent or ungrammatical output" (Yang). The compromise, a combination extractive-abstractive approach, allows for high-quality summaries that are still original text written by the model.

Sentiment Model

For our sentiment analysis model, we created a word-sentiment classifier based heavily on a tutorial by Robyn Speer of ConceptNet, an open-source semantic network used to create word embeddings (2017). The sentiment model is quite simple as it uses a logistic loss function trained on static word embeddings provided by ConceptNet and a word sentiment lexicon from Liu and Hu (2004). The benefit of using a logistic regression model was the speed at which we could train and use the model. Its accuracy at classifying the words in the test as positive or negative was greater than 97%, which gave us confidence in its ability to predict sentiment for word embeddings outside of the sentiment lexicon.

Data

We used articles from the CNN/Dailymail dataset as our summarization input text. The dataset is available from TensorFlow Datasets with examples already partitioned into Training Validation, and Test splits. After processing the text to include sentence separator tokens, we ran BertExtAbs on the text of articles from the first 5300 examples of the validation data, resulting in 5300 pairs of articles and generated summaries for us to use for analysis. The choice of 5300 examples was arbitrary and informed by availability of time and computing resources.

3 Results and Discussion

Sentiment Score Definition

Our analysis revolves around the difference in "sentiment score" between the input articles and the output summaries generated by BertExtAbs. We may sometimes refer to this difference in sentiment scores as the "change in sentiment score." The sentiment score for a word is generated by taking the difference in the log probability of a positive classification and the log probability of a negative classification. Thus, for a change of +1 in the sentiment score, the model is twice as likely to classify that word as positive.

To get the sentiment score for a text passage of more than one word, the sentiment scores of every known word in the input text is averaged, while words outside of the word embedding vocabulary are ignored. While this method is crude, it is adequate for our purposes of investigating the summarizing model's behavior as a whole. One consequence of our sentiment model choice is the caveat that any sentiment score we observe is most accurately interpreted as a description of word-choice sentiment rather than a classification of the sentiment of an input's semantic features.

Sentiment Change from Article to Summary

A histogram of the sentiment scores of all 5300 articles and summaries are displayed in Figure 1. We found that the distribution of sentiment scores for summaries was shifted slightly negative compared to that of the article sentiments. It was also relatively flatter, with a wider range of sentiment scores given to the summaries. This result is not quite informative about BertExtAbs because any effects the summarization model had on sentiment score is indistinguishable from effects that were driven by idiosyncrasies of the sentiment model. For one, the average summary word count is about one tenth that of the average article. A likely consequence of this is that there is more variance in the sentiment scores of summaries, since a passage's sentiment score is just the mean of those of known words.

In a similar but more speculative vein, the negative shift in sentiment scores may be a consequence of our choice of sentiment lexicon and word embeddings rather than the summarizing model's word choices. The sentiment lexicon contained about twice as many negative words as positive which could cause differences in the sentiment classifier's abilities to classify negative and positive words. The vocabulary of our chosen word embeddings may not be evenly distributed between positively and negatively classified words that appear in the text we are analyzing.

Examining the distribution of the change in sentiment scores (summary sentiment score minus article sentiment score) yields an approximately normal distribution, with most summary sentiment scores being fairly similar to and slightly lower than that of their corresponding articles (figure 2). We had hoped to find patterns of similarities between article/summary pairs that were outliers (i.e. having highly different sentiment scores between them), but we failed to detect any. Further investigation may be warranted.

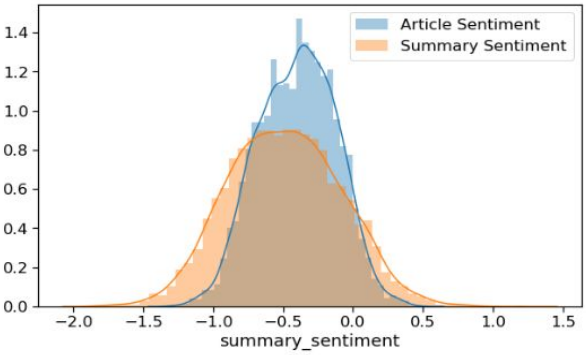

Figure 1: Distribution of sentiment scores for all 5300 articles and summaries.

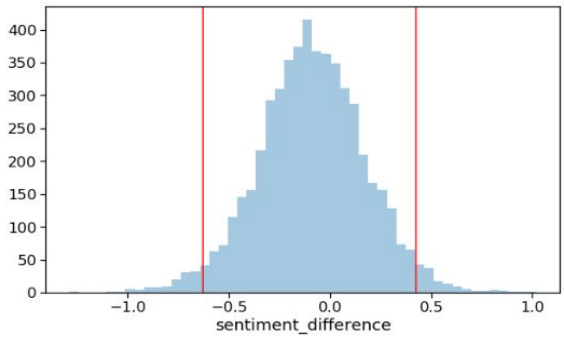

Figure 2: Distribution of Change in Sentiment, or the difference in sentiment observed between paired input and output passages.

Grouping Data By Gender

To explore the summarization model's potential biases with regards to gender, we sorted the article/summary pairs into two groups: examples about men and examples about women. To sort the examples, we assigned every passage a value we name its "gender coefficient." To calculate the gender coefficient of any passage, we subtract the count of male pronouns (he/him/his) from the count of female pronouns (she/her/hers) and divide this difference by the total count of pronouns in the passage, normalizing it. What results are gender coefficients of -1.0 for examples containing only male pronouns and +1.0 for those containing only female pronouns, with values in between for passages containing mixed gendered pronouns. We refer to these passages as "male-subject" or "female-subject" text, respectively.

```
female-subject articles count: 704
female-subject summaries count: 705
number in agreement: 704

male-subject articles count: 1879
male-subject summaries count: 1878
number in agreement: 1878
```

Figure 3: Counts of female- and male-subject examples. They have gender coefficients of 1.0 and -1.0, respectively.

Our data showed that the summarization model preserved the gender coefficient nearly perfectly, one indicator of the power of the BERT-based summarization model (Figure 3).

For simplicity, we chose to use in our analysis only examples whose original articles were exclusively male- or female-subject. Of our 5300 original examples, 1879 articles were male-subject and 704 were female-subject, meaning that there were nearly three times as many articles exclusively about men as there were of women in our data (Figure 4). As an aside, it's possible that this inequitable proportion is due to bias in the way the data were sampled for the CNN/Dailymail dataset, but we believe it is more likely to be caused by systemic sexism in journalism and our society as a whole.

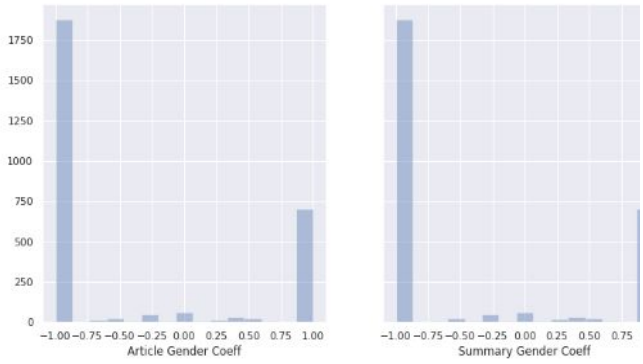

Figure 4: Counts of articles (left) and summaries (right) with varying gender coefficients.

Differences in Sentiment Change by Gender

We based our analysis of gender bias on the differences in change in sentiment between the female- and male-subject examples. The reason for this is that there are many ways in which gender bias that did not originate in the summarization model could affect absolute sentiment scores. For example, more articles about women may be more negative due to the stories that journalists choose to report. And our sentiment model itself may have inherited some bias from the data on which we trained it. Instead, comparing the relative changes in sentiment when grouping examples by gender would allow us to isolate and observe the bias coming solely from the summarization model's word choices. A large difference in the change in sentiment between female- and male-subject examples would be an indicator of a large gender bias.

We repeated the examination of sentiment score distributions of articles and summaries for our two groups of gendered articles. The results are similar, with a slightly wider and more negative distribution of scores for summaries. Between the female- and male-subject groups, the distributions are virtually indistinguishable visually (Figure 5)

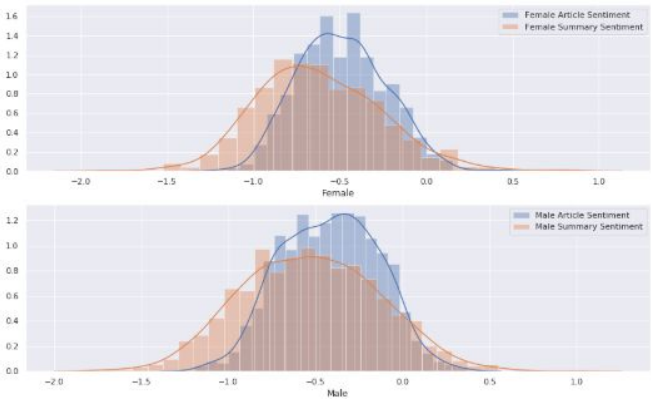

Figure 5. Distributions of sentiment scores of summaries and articles for female-subject (top) and male-subject (bottom) examples.

We performed an ordinary least squares regression of change in sentiment (from article to summary) on the input article sentiment score for both groups of article/summary pairs (Figure 6). We found that for female-subject articles, the change in sentiment from summarization does not change with article sentiment scores. However, for male-subject articles, the statistically significant coefficient of 0.0963 for the variable representing article sentiment score can be interpreted to mean that every increase of

1.0 in an article's sentiment score yields about a 0.1 increase in the corresponding summary's sentiment score. In other words, more positive articles result in summaries that are just slightly more positive when the article only references men.

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.0710	0.010	-6.765	0.000	-0.092	-0.050
article_sentiment	0.0963	0.021	4.598	0.000	0.055	0.137

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.1177	0.021	-5.660	0.000	-0.159	-0.077
article_sentiment	0.0273	0.038	0.722	0.471	-0.047	0.102

Figure 6. Regression summary for regression of change in sentiment on article sentiment, grouped by male-subject only (top) and female-subject only (bottom).

We followed this by running another linear regression with change in sentiment as the dependent variable and a dummy variable representing the subject-gender of articles (male=1, female=0) to compare the average change in sentiment (Figure 7.). The results showed that, on average, male-subject articles resulted in summaries that were scored 0.0198 points more positively than that of female-subject articles. However, with a p-value of 0.075, this estimate does not quite achieve statistical significance. Additionally, the effect size observed, if it reflects the true value resulting from gender bias, is very small.

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.1311	0.010	-13.788	0.000	-0.150	-0.112
is_male	0.0198	0.011	1.780	0.075	-0.002	0.042

Figure 7: Summary for regression of change in sentiment on is_male dummy variable. Only examples with -1.0 or +1.0 gender coefficients included.

By examining the difference in the differences of sentiment scores of text input/output pairs, grouped by gender, we hoped to mitigate the influence of any sources of bias outside of the summarization model we used. In turn, we can interpret the biases found in the outputs of the summarization model as having originated in the pre-trained BERT language model.

Our results show that, in the abstractive summarization setting, BERT is fairly unbiased but not yet perfect. While our approach treated BERT as a black box, favoring the examination of real inputs and outputs of a fine-tuned model, there is active research being done in identifying and mitigating bias in less superficial manners (Webster et al., 2018, 2019; Kurita et al. 2019).

Future iterations of this work would be greatly enhanced by the inclusion of many more input/output examples to analyze. In addition, it should not be difficult to apply more sophisticated forms of sentiment analysis using deep learning approaches in a similar framework to ours, allowing for semantic sentiment classification. Research with approaches similar to ours may be part of future toolkits for testing if attempts to de-bias language models are successful.

4 Conclusion

We found that sentiment classification is less precise on shorter text summaries than it is on corresponding articles. We found that change in sentiment scores, while only minutely biased, still shows a bias against women. Further research using this input/output comparison framework should seek to use larger sets of data for improved statistical power.

Sources

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. In CoRR, abs/1704.01444, 2017.

Matthew Peters, Mark Neumann, Mohit Iyyer,Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP).*

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018a. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 1797–1807*, Brussels, Belgium.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems,* pages 1693– 1701.

Evan Sandhaus. 2008. The New York Times Annotated Corpus. *Linguistic Data Consortium, Philadelphia,* 6(12).

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems,* NIPS'16, pages 4356–4364, USA. Curran Associates Inc.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the GAP: A Balanced Corpus of Gendered Ambiguous Pronouns. In *Transactions of the Association for Computational Linguistics, pages 605-617*.

Kellie Webster, Marta R. Costa-jussa, Christian Hardmeier, and Will Radford. 2019. Gendered ambiguous pronoun (GAP) shared task at the gender bias in NLP workshop 2019. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing, pages 1–7,* Florence, Italy. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointergenerator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* pages 1073– 1083, Vancouver, Canada.

Hu M., Liu B. Mining and summarizing customer reviews. *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '04)*; August 2004; pp. 168–177.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations