Transforming Text with Transformer Models: Exploring BERT and T5's Potential in Bias-Free NLP Applications

Ramsha Perwez
College of Computing & Informatics
Drexel University
Philadelphia, PA 19104, USA
rp955@drexel.edu

Abstract

Language learning models are powerful tools that can generate intelligent text with the guidance of fine-tuning. This project harnesses a machine learning pipeline for detecting and mitigating bias in text. It employs multiple models such as logistic regression and BERT to analyze text for bias and generate alternative phrasings. Initial bias and toxicity measurement datasets were pulled in from the 'Jigsaw Unintended Bias in Toxicity Classification' [1] available on Kaggle. The trained portion of this dataset has almost two million records. Data preprocessing and feature extraction takes place using TF-IDF vectorization and pushing through a logistic regression classifier to detect biased language. To generate alternative text, T5 Text-to-Text Transformer Model and BERT-based models are experimented for phrasing suggestions. The project demonstrates the practical application of natural language processing techniques to address the important issue of bias in communication.

1 Introduction

In modern society, there is an ever-growing need to mitigate bias, even in communications stemming from the most neutral-minded intentions. As our world becomes increasingly interconnected and diverse, the language we use plays a crucial role in shaping perceptions, influencing decisions, and fostering inclusivity. This era is the perfect playground for expanding research in linguistic training. There are multiple limitations that should be acknowledged that this report will acknowledge along the way: data bias, context-sensitivity and the biggest hindrance in this experiment: computational resources. Nevertheless, this start point provides a great start point for...

2 All About the Data

The key to a robust experiment in natural language processing and content moderation is an abundance of data, which the Jigsaw Unintended Bias in Toxicity Classification dataset provides. Although not intended for explicit bias, this crucial resource was designed to detect toxic comments and prevent unintended microaggressions. The `toxicity` feature and `target` label offer a nuanced scoring system (0 to 1) for toxicity levels, while additional attributes like `severe_toxicity`, `obscene`, and `identity_attack` enable a multifaceted approach to content analysis. The dataset's inclusion of various identity categories (e.g., `asian`, `female`, `muslim`) examines the bias across demographic groups. With almost two million labeled comments, it provides the statistical power necessary for training machine learning models, capturing diverse manifestations of toxic language and bias.

3 What Time is it?

This experiment was performed on Google Colab which has a system RAM memory of approximately 12.7 GB. Various models such as Random Forests and Gradient Boosting were attempted; however, there were constant limitations encountered due to system RAM constraints and time restrictions. This led to decision making between model complexity and computational feasibility. The more complex models, while potentially more accurate, often exceeded the available memory or took an impractically long time to train. As a result, simpler, more efficient algorithms were prioritized that could run within the given constraints while still providing reasonable performance.

4 Model Training Process for Bias Detection

The model training segment begins with loading data from a CSV file (API is available on Kaggle) and preprocessing features for training. The model is trained with the scikit-learn pipeline, combining TF-IDF vectorization and logistic regression. TF-IDF vectorization paired with logistic regression gave the most promising statistical results: an accuracy of 0.89, and the best recall of 0.81 from the available data. It suggests that this model is strong at identifying potential bias. However, the low precision indicates that overall, there is significant room for eliminating false positives. The initial statistical results serve as a great start and give room for experimenting with different models for improving overall performance.

4.1 Model Evaluation Metrics

Metric	**Value**
Accuracy	0.8938
Precision	0.4156
Recall	0.8146
F1 Score	0.5504

Table 1: Model evaluation metrics of for Bias Detection using Logistic Regression and TF-IDF Vectorizer

5 T5 Text-to-Text Transformer Model

Initial experimentation started with the T5 Text-to-Text Transfer Transformer. The implementation used the `debias_text` function, which employs T5 to rephrase potentially biased statements. The function begins with "rephrase this following text:", instructing T5 to generate an alternative version. The results of this approach were mixed. For the statement "Women are too emotional for leadership positions," T5 successfully generated a less biased alternative: "Women are more powerful for leadership positions," which the bias predictor classified as not biased. However, for other examples, the results were not as accurate. The debiased version of "Immigrants are stealing our jobs" became "people are killing our land," which was rated as even more biased than the original. Similarly, T5 struggled with the highly biased statement about women and high-pressure jobs, producing another biased statement. The temperature was reduced to 0.5 to balance creativity in the output. While T5 demonstrated an ability to rephrase and potentially reduce bias in some cases, it also showed a tendency to introduce new, unintended biases or fail to adequately address the original bias. There were attempts to experiment with varying the model size, increasing the number of training epochs, and adjusting output parameters such as temperature and sampling methods in the T5 model. Despite these changes, there were no improvements in the results, suggesting that further adjustments or a different approach might be required [2].

5.1 T5 Model Text Generations

Original Text	Immigrants are stealing our jobs.
Bias Prediction	Moderately biased (Score: 0.43)
Debiased Version	People are killing our land.
Bias Prediction of Debiased Version	Highly biased (Score: 0.78)

Table 2: Bias prediction for the original text and its debiased version.

Original Text	Women are too emotional for leadership positions.
Bias Prediction	Moderately biased (Score: 0.42)
Debiased Version	Women are more powerful for leadership positions.
Bias Prediction of Debiased Version	Not biased (Score: 0.25)

Table 3: Bias prediction for the original text and its debiased version.

Original Text	Women are too dumb to handle high-pressure jobs.
Bias Prediction	Highly biased (Score: 1.00)
Debiased Version	Women are too weak to accept their water.
Bias Prediction of Debiased Version	Highly biased (Score: 0.78)

Table 4: Bias prediction for the original text and its debiased version.

6 BERT Model

The subsequent approach leverages BERT's masked language modeling capabilities to address potential bias in the input text. The `debias_text` function processes each word in the input text by iteratively masking it and querying BERT's contextual understanding. For each masked token, BERT generates probabilistic predictions. If the original word falls outside BERT's top-k predictions (where k=5 in this implementation), the function initiates a swapping procedure, selecting an alternative from BERT's suggestions. This methodology effectively substitutes potentially biased terms with more neutral alternatives derived from BERT's language model, balancing between bias mitigation and preservation of original meaning [3][4].

6.1 Generation of Alternatives Using BERT

The `generate_alternatives` function begins by tokenizing the input text and tagging parts of speech, which involves iterating through the words and identifying their grammatical roles. The function then generates alternative phrasings by randomly masking and replacing words based on their grammatical roles. This involves multiple forward passes through the BERT model to predict replacements for the masked words. For example, in the sentence "Women are too emotional for leadership positions," "emotional" might be masked. The masked text is then processed by BERT, which predicts possible replacements for the [MASK]. This masking technique enables the exploration of different ways to express the same idea [4][5].

7 Generating Alternative Phrasings with Bias Predictions

Original Text	**Bias Prediction**
Immigrants are stealing our jobs.	Moderately biased (Score: 0.43)
Alternative 1	**Bias Prediction**
They are stealing our jobs.	Not biased (Score: 0.24)
Alternative 2	**Bias Prediction**
Immigrants are doing our jobs.	Not biased (Score: 0.23)
Alternative 3	**Bias Prediction**
Immigrants are doing our jobs.	Not biased (Score: 0.23)
Alternative 4	**Bias Prediction**
Immigrants are doing our jobs.	Not biased (Score: 0.23)
Alternative 5	**Bias Prediction**
Immigrants are doing our jobs.	Not biased (Score: 0.23)

Figure 1: Bias prediction for the original text and its alternatives.

Original Text	**Bias Prediction**
Women are too emotional for leadership positions.	Moderately biased (Score: 0.42)
Alternative 1	**Bias Prediction**
Women are often considered for leadership positions.	Not biased (Score: 0.13)
Alternative 2	**Bias Prediction**
They are very emotional for leadership positions.	Not biased (Score: 0.09)
Alternative 3	**Bias Prediction**
Women are emotional for leadership positions.	Moderately biased (Score: 0.44)
Alternative 4	**Bias Prediction**
Women were too young for leadership positions.	Not biased (Score: 0.11)
Alternative 5	**Bias Prediction**
Women are too young for these positions.	Highly biased (Score: 0.75)

Figure 2: Bias prediction for the original text and its alternatives.

Original Text	**Bias Prediction**
Women are too dumb to handle high-pressure jobs.	Highly biased (Score: 1.00)
Alternative 1	**Bias Prediction**
Women are too dumb to take high-pressure sex.	Highly biased (Score: 1.00)
Alternative 2	**Bias Prediction**
Women are often required to handle high-pressure jobs.	Not biased (Score: 0.08)
Alternative 3	**Bias Prediction**
They are too dumb to handle high-pressure water.	Highly biased (Score: 1.00)
Alternative 4	**Bias Prediction**
Women were too young to handle high-pressure jobs.	Not biased (Score: 0.06)
Alternative 5	**Bias Prediction**
They are too dumb to handle these jobs.	Highly biased (Score: 1.00)

Figure 3: Bias prediction for the original text and its alternatives.

7.1 Testing BERT Inference

Model inferencing can give classifications on unseen data. In the context of BERT-based bias detection, the model can apply itself to new unseen text to consider bias. There is no further learning involved, but rather patterns that were learned during initial training to make predictions or classifications. The results in this experiment indicated some uncertainty with the model's ability. When presented with the statement "Immigrants are stealing our jobs," the model correctly identified it as biased with 55.42% confidence. This classification reflects the model's ability to recognize negative stereotypes although with some uncertainty. Interestingly, when processing the neutral statement "The sky is blue," the model correctly identified it as unbiased, but with a very slight margin – 50.06% confidence. This near-equal split demonstrates the model's appropriate caution when faced with statements lacking clear bias indicators, showing a tendency to avoid false positives in bias detection. Although there is equal argument that the model should recognize fact versus fiction and thus give it a higher confidence score.

7.2 BERT Model Inference Probabilities

Input Text	**Predicted Class**	**Class Probabilities**
Immigrants are stealing our jobs.	1	[0.4458, 0.5542]
Women are too dumb to handle high-pressure jobs.	1	[0.4569, 0.5431]
The sky is blue.	0	[0.5006, 0.4994]

Figure 4: Model predictions and class probabilities for various input texts.

8 Further Considerations

The BERT model successfully identifies clear instances of bias, as seen in its consistent classification of statements like "Women are too dumb to handle high-pressure jobs" as highly biased. However, its performance on more nuanced statements reveals the complexity of bias detection. Future improvements should focus on increasing the model's confidence in clear-cut cases, fine-tuning its threshold for borderline cases, and expanding its training data to cover a wider range of subtle and explicit biases. This program utilized free models which are less powerful along with limited computational capacity. More advanced transformer architectures like GPT-3 or BART might provide enhanced context understanding and nuanced language interpretation. Multi-modal models that incorporate both text and visual data could provide a more comprehensive understanding of bias as well. For the power that it holds, this BERT model for bias detection demonstrates both promising capabilities and areas for improvement.

References

[1] "Jigsaw Unintended Bias in Toxicity Classification." *Kaggle*, www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification. Accessed 6 Sept. 2024.

[2] "T5." *Hugging Face*, huggingface.co/docs/transformers/en/model$_doc/t5.Accessed6Sept.$2024.

[3] Devlin, Jacob, et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *arXiv*, vol. 1, 2018, arXiv:1810.04805.

[4] Hugging Face. *BERT Model Documentation*. Hugging Face, https://huggingface.co/docs/transformers/en/model$_doc/bert.A$

[5] Radhakrishnan, S. "Natural Language Inference Using BERT and PyTorch." *Medium*, 10 Sept. 2020, https://medium.com/red-buffer/natural-language-inference-using-bert-and-pytorch-6ed8e69f93bc. Accessed 7 Sept. 2024.