Deep Learning Models for Subjective Bias Neutralization

Violet Davis
vdavis@berkeley.edu
University of California, Berkeley
Master of Information and Data Science

Abstract

Understanding and mitigating subjective biases are crucial for promoting fairness, equity, and inclusion, yet biased language remains pervasive in media, organizations, and even in current large language models. This project develops deep learning models for automated bias neutralization and classification. Text-to-text encoder decoder transformer models are trained to automatically neutralize biased language and measured using BLEU and accuracy scores. This paper also proposes methods of measuring subjective bias automatically through convolutional neural networks assessed using accuracy, precision, recall, and F1-scores. The high performance of a fine-tuned large language model indicates ample opportunity for their implementation in nuanced downstream tasks regarding bias in text, while the classification models demonstrate the potential of automated bias measurement.

I. Introduction

In recent years, researchers have made significant progress in identifying and addressing bias in toxic language and distinct demographic categories such as gender and race. However, the detection and neutralization of more nuanced, intersectional bias remains an emerging field. Subjective bias, which stems from personal opinion instead of fact, poses a great challenge for machine learning algorithms to detect. This type of bias pervades across domains and marginalized groups, yet is often not as explicit as overt hate speech. Nevertheless, it still can inflict significant harm on minority groups when the personal opinions reflect underlying biases regarding race, religion, gender, and sexuality.

For many individuals, these biases may be unconscious or subconscious. For instance, many people who do not perceive themselves as homophobic may unknowingly employ heteronormative language that can be exclusionary. The development of tools capable of automatically identifying and neutralizing biased text would be extremely beneficial to people and organizations who strive to communicate with unbiased language.

Additionally, large language models are increasingly utilized in natural language processing and by non-technical people in various domains, yet even these models can, and often do, contain subjective bias. Further work, such as fine-tuning to address and rectify the models' biases, is critical to ensure equitable and inclusive language is used by all.

II. Related Work

Numerous researchers have addressed bias in natural language processing by directly modifying word embeddings. The paper "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings" (Bolukbasi et al. 2016) became a prominent example of addressing gender bias in embeddings. In their work, word vectors are transformed to mitigate gender stereotypes, with approaches including the removal of biased gender associations and the promotion of gender neutrality through equalization techniques.

In the years following, with the introduction of increasingly large language models, NLP researchers have shifted their focus towards fine-tuning these models. Many have developed corpuses for fine-tuning related to gender, race, religion, heteronormativity (Vásquez et al. 2022), etc. One such dataset, The Wiki Neutrality Corpus (Pryzant et al. 2019), accumulated a parallel corpus of subjectively biased and neutralized sentences. The accompanying paper “Automatically Neutralizing Subjective Bias in Text” proposes modular and concurrent models to generate neutral text from biased text, and classifies the generated text for bias, fluency, and meaning metrics. However, it's worth noting that these metrics are evaluated using thousands of crowdworkers, not assessed automatically.

Automatic calculation of nuanced bias metrics in natural language processing (NLP) has presented challenges. Commonly used resources include WEAT (Caliskan et al. 2017), SEAT (May et al. 2019) and CEAT (Guo et al. 2020), which measure implicit embedding bias associated with specific attribute groups, such as gender. Additionally, CrowS-Pairs (Nangia et al. 2020) and StereoSet (Nadeem et al. 2021) are often utilized to measure stereotype bias for masked language models. For toxic and hate speech, Perspective API[1] is a free API made by Jigsaw and Google’s Counter Abuse Technology team to measure toxicity in text. Initially, the tool received negative press for mislabeling statements such as “I am gay” as toxic, but the team has since been refining it with the help of human-generated feedback. Ongoing research developments continuously expand the resources available for those attempting to address and mitigate bias in NLP systems automatically.

III. Data

The data comes from the Wiki Neutrality Corpus (WNC) (Pryzant et al. 2019), a parallel corpus of over 180,000 sentence pairs categorized into biased and neutralized versions. The corpus was curated by scraping Wikipedia edits made under the platform's "Neutral Point of View" policy, which encourages users to amend sentences to adhere to neutral, factual, and nonjudgmental language standards.[2] The WNC includes sentences related to three primary forms of subjective bias: epistemological bias, framing bias, and demographic bias.

The WNC data was used to train all of the models, both text-to-text generation and bias classification. The data was split into 127,033 training pairs, 27,220 validation pairs, and 27,220 test pairs. For bias classification models, the data sizes doubled as each pair was split and labeled, with “0” for the biased source sentence and “1” for the neutralized target sentence.

[1] https://perspectiveapi.com/
[2] https://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view

IV. Methods

Text-to-Text Generation

As a baseline, we trained a sequence-to-sequence encoder decoder transformer model. Our approach was to consider the task as a machine translation task, with different learned vocabularies and tokenizers for biased and neutral language.

To improve this model, we fine-tuned the T5 base model on the WNC data. With computational constraints, we trained for only one epoch with a maximum token length of 128 on a NVIDIA A100 GPU. (Ideally, we would use a maximum token length of 256 or 512 to capture even the longest sentences, but would need substantial computational resources.) We selected the T5 base model for its encoder decoder architecture, vast pre-trained corpus of data and transfer learning capabilities.

Bias Classification

As a baseline, we trained a BiLSTM CNN model with attention. This architecture was adapted from code successfully implemented for toxic comment multi-class classification[3]. The model features a single convolutional layer with 64 filters and kernel size 3. By incorporating a combination of BiLSTM, CNN, and attention mechanisms, we aimed to capture contextual nuances and patterns within sentences. Given that our data size doubled as each sentence pair now became two individual data points, we experimented with training data selection strategies. In the initial approach, we randomly selected either the source or target sentence from each sentence pair in all datasets. In the subsequent model iteration, we included both source and target sentences, but limited the model to only using 130,000 of the total 254,066 training sentences due to computational limitations.

To improve this model, we trained a BERT binary classification CNN model. We hypothesized that the addition of the pre-trained BERT base and transformer architecture would boost performance over our baseline model. The architecture employs three convolutional layers with 64 filters each and kernel sizes of 3, 5, and 7. After running the two BiLSTM models, the second method of data selection reduced overfitting and slightly increased metrics, so we only implemented this method for the BERT CNN. With resource limitations, training was restricted to three epochs using a NVIDIA A100 GPU and convergence was not yet reached.

V. Results

Table 1: Example Source, Target, and Generated Test Sentences

WNC Source "Biased" Sentence	**WNC Target "Neutralized" Sentence**	**Seq2Seq Generated Sentence**	**T5 Fine-Tuned Generated Sentence**
The player must not make any move that would place his king in check.	The player must not make any move that would place their king in check.	The player must not make any move that would place in his .	The player must not make any move that would place their king in check.

[3] https://www.kaggle.com/code/tranduongminhdai/attention-based-bilstm-cnn-approach

The lyrics are about mankind's perceived idea of hell.	The lyrics are about humanity's perceived idea of hell.	The lyrics are about humankind's perceived idea of hell.	The lyrics are about humankind's perceived idea of hell.
Marriage is a holy union of individuals.	Marriage is a personal union of individuals.	Marriage is a holy union of individuals.	Marriage is a union of individuals.

As seen in Table 1, the T5 fine-tuned model effectively mitigates bias and neutralizes language. Although it does not always output the exact wording of the target sentence, it generates sentences with similar meaning. On the contrary, the sequence-to-sequence model is inconsistent. In some examples, it doesn't change the original sentence, in some cases it does neutralize the language, and in others still, it generates grammatical errors.

Table 2: Text-to-Text Generation Model Analysis

Text-to-Text Generation Model	**BLEU Score (Average)**	**BLEU Score (Overall)**	**Exact Accuracy**
Sequence-to-Sequence	66.95%	71.08%	00.06%
T5 Fine-Tuned	87.83%	91.93%	25.87%

For each text-to-text generation model, we include three metrics of evaluation in Table 2 as well as an additional bias metric in Table 4. In the initial analysis, we evaluate average BLEU score, overall BLEU score and Exact Accuracy, measured by the percentage of exact matches to the target sentences.

The performance of the sequence-to-sequence model appeared strong at first with high accuracy during training. However, upon closer investigation, the model predominantly generates the same sentence as the input. This outcome could be caused by our usage of a subword model when the sentence pairs have a substantial overlap in vocabulary. As well, since each pair of sentences has only minor, nuanced changes, if the model generates the source sentence, it still appears to be a good match to the target. Therefore, it is no surprise that the BLEU scores are low for our task of bias neutralization, but still over 50%. The Exact Accuracy of the model of 0.06% aligns with the example sentences in Table 2. In the one case where the model neutralized the language from "mankind" to "humankind", the target of "humanity" was still not matched.

The T5 fine-tuned model performs significantly better with an Overall BLEU Score comparable to the top-performing models proposed by Pryzant et al. This demonstrates that fine-tuning large language models can mirror performance of much more complicated model architectures. The Exact Accuracy score of the model aligns more closely to the Base or Transformer models in the paper. This discrepancy likely stems from our model leveraging transfer learning and thus, drawing upon phrasing variations from pre-trained text and tasks. As evident in the test examples, the T5 model consistently generates neutralized sentences conveying the same meaning but with different wording than the target sentences.

Table 3: Classification Model Analysis

Classification Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
BiLSTM CNN (Random Choice)	58.97%	58.66%	61.64%	60.11%
BiLSTM CNN (130,000 Training Sentences)	60.75%	60.37%	62.56%	61.45%
BERT CNN (130,000 Training Sentences)	69.13%	66.02%	78.81%	71.85%

For each classification model, we analyze Accuracy, Precision, Recall, and F1-Score. True positives represent correctly identified neutralized sentences and true negatives represent correctly identified biased sentences. Recall emerged as the highest-performing metric across all models. Given the parity between positive and negative classes, this indicates that the models are better at identifying neutral sentences compared to identifying biased sentences.

The BiLSTM CNN model with random choice sentence selection quickly overfit the training dataset, which is why we did not use this training data selection method for the BERT CNN model. The improved performance of the BERT CNN may largely stem from its architecture, which incorporates 3 convolutional layers instead of just one. However, it is worth noting that while we trained the BiLSTM models for ten epochs each, we trained the BERT CNN model for only three epochs due to the significantly increased computational demands. Ideally, we would have trained this model for ten epochs as well to see if it would reach convergence and yield superior results.

Table 4: Source, Target, and Generated Test Sentences Analysis

	WNC Source "Biased" Sentences	**WNC Target "Neutralized" Sentences**	**Seq2Seq Generated Sentences**	**T5 Generated Sentences**
Average BERT CNN Bias Score	60.54%	37.62%	42.92%	39.12%

The average BERT CNN Bias Scores are calculated by averaging the predicted bias label percentage for each sentence. For reference, scores above 50% are classified as "biased" while scores below 50% are categorized as "neutral." As expected, the source sentences average above 50% whereas the target sentences average below 50%, though by only 11 percentage points above and 12 points below respectively. The bias scores of the generated sentences give us further information on the effectiveness of the models. Surprisingly, the scores of the two models are remarkably similar with only three percentage points separating them. The T5 fine-tuned model exhibits slightly less biased results, with an average score very near to that of the target sentence corpus. The sequence-to-sequence model trails slightly behind with a marginally higher score, albeit still averaging within the "neutral" territory.

VI. Conclusion

In this project, we established a framework for implementing both automated text neutralization and bias metrics. For text-to-text generation, the sequence-to-sequence transformer model demonstrated variable performance in generating neutralized text, while the fine-tuned T5 model exhibits notable proficiency, approaching the performance levels of previous studies in just one epoch. For bias classification, we encountered the challenge of training an abundance of data with limited computational resources. The BERT CNN model shows strong promise and improved performance over the BiLSTM CNN models, but would require significant computational power to reach convergence. However, improving the quality of the bias measurement metric with additional training could prove instrumental for further research or mainstream use.

Future work could include applying these frameworks to novel datasets to measure bias and generate neutralized text, as well as further analyze the models for generalizability across domains. Additional applications could include neutralizing language in speeches or articles. Different linguistic styles compared to the training data of Wikipedia articles could impact performance and would likely further widen the gap between the performance of large language models and architectures with limited learned vocabularies.

A further extension of this project would be developing automated bias neutralization and bias measuring tools for real world applications. The need for such solutions is substantial, given the pervasive nature of biases and the growing recognition of the need to address them effectively. The development of an application or chatbot using these models could offer practical solutions for mitigating biases in contexts such as social media or news. By automating bias detection and neutralization, these tools could help promote the usage of more inclusive, unbiased language, potentially leading to broader positive impacts across society.

References

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings.

Aylin Caliskan, Joanna J Bryson, Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, Rahul Gupta. 2021. BOLD: Dataset and Metrics for Measuring Biases in Open-Ended Language Generation

Michael Gira, Ruisu Zhang, and Kangwook Lee. 2022. Debiasing Pre-Trained Language Models via Efficient Fine-Tuning.

Wei Guo, Aylin Caliskan. 2020. Detecting Emergent Intersectional Biases: Contextualized Word Embeddings Contain a Distribution of Human-like Biases

Yue Guo, Yi Yang, Ahmed Abbasi. 2022. Auto-Debias: Debiasing Masked Language Models with Automated Biased Prompts.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, Rachel Rudinger. 2019. On Measuring Social Biases in Sentence Encoders

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models.

Flavien Prost, Nithum Thain, Tolga Bolukbasi. 2019. Debiasing Embeddings for Reduced Gender Bias in Text Classification

Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, Diyi Yang. 2019. Automatically Neutralizing Subjective Bias in Text

Yolande Strengers, Lizhen Qu, Qiongkai Xu, Jarrod Knibbe. 2020. Adhering, Steering, and Queering: Treatment of Gender in Natural Language Generation.

Juan Vásquez, Gemma Bel-Enguix, Scott Thomas Andersen, Sergio-Luis Ojeda-Trueba. 2022. HeteroCorpus: A Corpus for Heteronormative Language Detection.