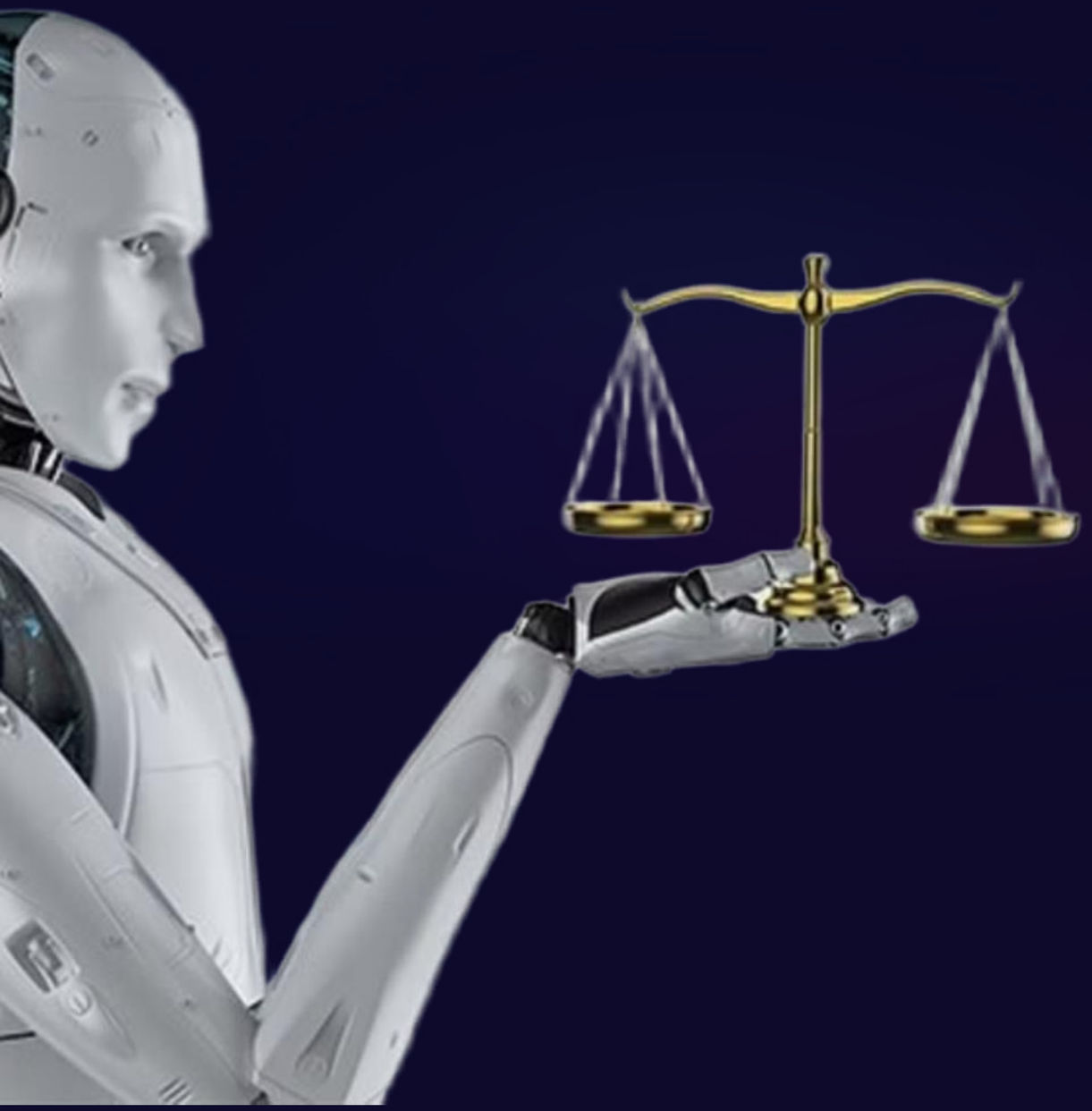


Modular Pipeline for Gender Bias Migration in Short Texts

:Names

Katherine Zablianov
Ariel Soffer
Shy Yeffet
Neta Robinzon Butbul

Modular Pipeline for Gender Bias Mitigation in Short Texts

- **Detect gender bias** in text.
- **Rewrite biased sentences** to neutral ones without changing the meaning.

Tasks + Data:

Classifier Model:

- Model: bert-base-uncased model

Fine-Tuned on: MGSD (Merged Gender Stereotype Dataset) - consisting labeled sentences (gender biased/ neutral)

- Output: Bias score (range 0–1)
- Threshold: 0.5 (for classification as biased)
- Evaluation: f1_score, accuracy, precision, recall metrics, tested on test_df from MGSD

Rewriting Method 1 – IG + Pretrained:

- Model: uses FLAN-T5 model for rewriting
- Tool: uses Integrated Gradients to identify top-k biased tokens
- Input: gender biased sentences that were classified as biased (in our classifier)
- Output: rewritten version of the original sentence with less (hopefully without) gender bias

Modular Pipeline for Gender Bias Mitigation in Short Texts

- **Rewriting Method 2 — Fine-Tuned Rewriter:**
- Model: T5 (Text-To-Text Transfer Transformer)
- Fine-Tuned on: winobias_gold_pairs - pairs of gender-biased sentences and their unbiased versions (biased, neutral)
- Output: rewritten version of the original sentence with less (hopefully without) gender bias

Strategy and step-by step plan:

Entering a sentence into the pipeline •
The sentence is classifies to biased/not biased •
If it is classified into biassed it goes through our 2 methods of rewriting •
The two rewritten sentences are then presented as an output •

Remark: We will preprocess all of our datasets, make sure there are no nulls, plot distribution and make sure the data is balamced, extract only the relevant columns and so

Once pipeline is trained and evaluated, it can be integrated with GPT models or **used as a smart pre/post-processing layer**

Evaluation – bias reduction and meaning preservation

Evaluate bias reduction in both rewriting models:

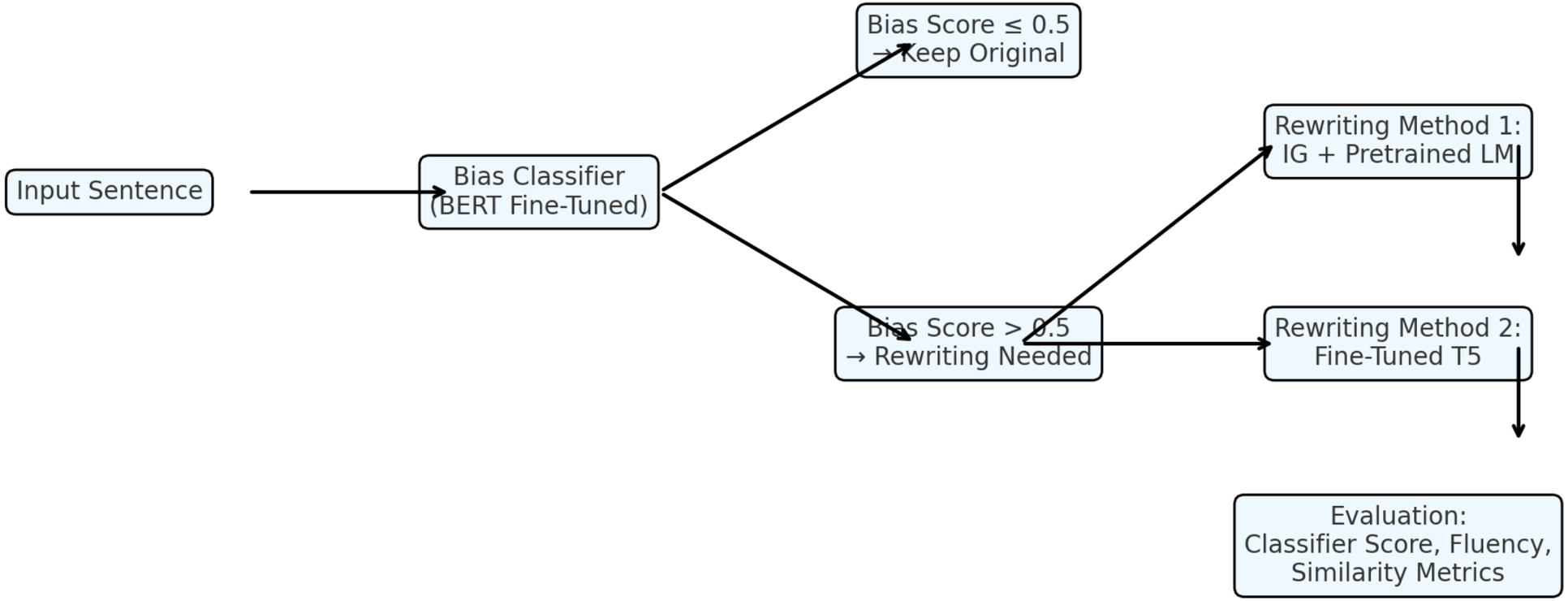
Using **SEAT (**Sentence-level association test) and **WEAT (**Word embedding association test) on outputs to assess whether **latent gender associations** are reduced
Using our bias classifier again with the rewritten sentences to check if the bias was reduced

Evaluation of the rewriting models - Meaning Preservation:

First Rewriting Model evaluation (IG + Pretrained Model):
Blacklist N-gram removal: Are flagged biased words removed
Cosine Similarity between original and rewritten embeddings

Second Rewriting Model evaluation (Fine-Tuned Rewriter):
Rouge and BERTScore metrics (paired biased vs. neutral reference)
Cosine Similarity between original and rewritten embeddings

We will also calculate a fluency Score using GPT perplexity



Previous Work

Source / Title	Approach / Model	Data	Metrics	Results
Dbias: Detecting Biases and Ensuring Fairness in News Articles (2022)	Two-stage pipeline: BERT-based classifier → GPT-2 rewriter with reward optimization	BiasFinder + custom news data	Classifier bias score, human judgment	Reduced bias scores by ~20%, and human judges preferred rewritten versions 79% of the time
From ‘Showgirls’ to ‘Performers’: Fine-tuning with Gender-inclusive Language (2024)	Fine-tuned LLMs with gender-neutral term replacements	Gender-transformed corpora + CrowS-Pairs	Stereotype score (CrowS-Pairs, StereoSet)	Reduced stereotype scores by 11% (RoBERTa) and 15% (GPT-2), with minimal loss in fluency
An Empirical Analysis of Parameter-Efficient Methods for Debiasing Pretrained LMs (2023)	Adapter, prompt, and LoRA tuning on BERT and GPT-2	CrowS-Pairs + StereoSet	Stereotype score, perplexity	Adapter tuning lowered stereotype scores by up to 18% with <1% increase in perplexity (maintained fluency)

Baselines

Baseline for classification:

Dataset:

150 sentences sampled from the MGSD dataset

Each sentence manually labeled as biased or neutral

Focus on gender-related bias only

Sentences vary in length and grammatical structure

Baseline:

Use a dictionary of gendered and stereotypical keywords

Classify sentence as biased if any keyword is matched

Compute a bias score as the ratio of matched keywords to total tokens

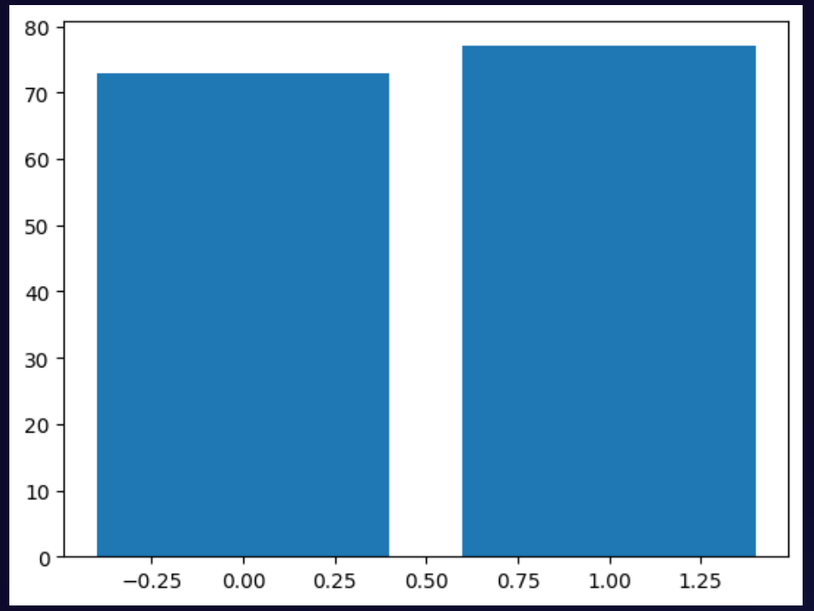
Evaluate on full 150-sentence set

Report accuracy, precision, recall, and F1 score

Evaluation Metrics:

Accuracy : 0.49 Precision: 0.50

Recall : 0.57 F1 Score : 0.53



	bias_score
0	0.0000
1	0.0000
2	0.2500
3	0.0000
4	0.1250
..	...
145	0.0000
146	0.1250
147	0.0000
148	0.2500
149	0.0625

Baselines

Baseline for rewriting model:

Dataset:

150 sentence pairs from winobias_gold_pairs.csv

Each example contains a biased sentence and its manually crafted **neutral rewrite**

Split: 100 examples for training, 50 for testing

Sentence lengths vary

Step	Training Loss
10	9.015600

```
Inference on device: 0
Device set to use cuda:0
```

	ROUGE-1	ROUGE-2	ROUGE-L
0	0.402547	0.35159	0.404245

```
C:\Users\sy020\.conda\envs\env_tran
ention. (Triggered internally at ..
  attn_output = torch.nn.functional
Average Cosine Similarity: 0.4851
Some weights of RobertaModel were n
berta.pooler.dense.weight']
You should probably TRAIN this mode
Average BERTScore F1: 0.4831
```

Baseline:

Fine-tune T5-small on 100 biased → neutral examples, trained on 1 epoch

Format inputs as: "neutralize: [biased sentence]"

testing is done on the first 50 held-out examples

Output: rewritten neutral sentences generated from biased inputs

Evaluation metrics we used in the baseline are Rouge (1, 2, L) for measuring **word-level and phrase-level similarity**, cosine similarity to measure **semantic similarity** between the **original biased sentence** and its **rewritten neutral version, and BertScore to** To evaluate **semantic overlap at the token level** using BERT embeddings (it validates that the rewrite **still means the same thing)**

Insights

The dictionary- based classifier method for classification provided relatively low accuracy (accuracy = 0.49, f1_score = 0.53), indicating that the dictionary was too general

we need to find more specific words that affect the bias in each sentence and capture specific nuances of bias

The bias score represents the ratio of biased words from the dictionary to the number of words in the sentence. It is very low in most cases

Therefore, it does not truly reflect the level of gender bias and we will change the way it is calculated so that it can measure the true level of bias

The model shows moderate surface similarity to references (ROUGE-1: 0.40, ROUGE-2: 0.35)

Therefore, we will train on a larger dataset (as expected) to improve pattern learning and structural fluency

The semantic preservation is weak (Cosine: 0.48, BERTScore F1: 0.48) — rewrites often drift in meaning

So we will Add semantic-aware training (e.g., embedding-based loss) and use guided editing with IG