Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24)
Special Track on AI for Good

From Pink and Blue to a Rainbow Hue! Defying Gender Bias Through Gender Neutralizing Text Transformations

Gopendra Vikram Singh[1†], **Soumitra Ghosh**[2†], **Neil Dcruze**[3*] and **Asif Ekbal**[4]
[1]Department of Computer Science and Engineering, Indian Institute of Technology Patna, Patna, India
[2]NLP Research Group, Fondazione Bruno Kessler (FBK), Trento, Italy
[3]Middlebury College, Middlebury, Vermont, United States
[4]School of AI and Data Science, IIT Jodhpur
{gopendra.99, neildcruze}@gmail.com, sghosh@fbk.eu, asif@iitj.ac.in

Abstract

In an era where language biases contribute to societal inequalities, this research focuses on gender bias in textual data, with profound implications for promoting inclusivity and equity, aligning with United Nations Sustainable Development Goals (SDGs) and upholding the principle of Leave No One Behind (LNOB). Leveraging advances in artificial intelligence, the study introduces the *GEnder-NEutralizing Text Transformation (GENETT)* framework, addressing gender bias in text through auto-encoders, vector quantization, and Neutrality-Infused Stylization. Furthermore, we present the first-of-its-kind corpus of *GEnder Neutralized REvisions (GENRE)* crafted from gender-stereotyped versions. This corpus serves a multifaceted utility, offering a resource for diverse downstream tasks in gender-bias analysis. Extensive experimentation on *GENRE* highlights the superiority of the proposed model over established baselines and state-of-the-art methods. Access the code and dataset at 1. https://www.iitp.ac.in/~ai-nlp-ml/resources.html#GNR, 2. https://github.com/Soumitra816/GNR.

Note: *Our research focuses on understanding cyber harassment conversations, especially in under-researched areas, with the exclusion of non-binary cases due to existing dataset limitations, not lack of sensitivity. We strive for inclusivity and plan to address this in future research with suitable datasets.*

1 Introduction

Recent years have seen increasing recognition of the biases and stereotypes in language that perpetuate societal inequalities and marginalize certain groups [Singh *et al.*, 2023]. Gender bias in textual data is one such form of bias that has far-reaching implications. Addressing and mitigating gender bias is crucial for promoting inclusivity and equality. With advancements in artificial intelligence (AI), specifically natural language processing (NLP), there is an opportunity to develop automated approaches to generate gender-neutralized versions of biased text, thus contributing to social impact and creating a more inclusive online space.

Type	Text
GB	You want equality? Learn to not pull the "I'm a woman" card.
GNR	If you seek equality, let go of any exclusive identity claims.
GB	Women can't park.
GNR	Some individuals struggle with parking.
GB	Why are the women in accounting so bitchy?
GNR	Why do some individuals in accounting have a negative attitude?
GB	Chicks just aren't that funny.
GNR	Some people just aren't that funny.
GB	Women's sexual desires are less intense than men's.
GNR	People's sexual desires can vary in intensity.

Table 1: Sample instances from the *GENRE* dataset. GB: Gender-biased, GNR: Gender-neutralized revision.

While existing studies have explored gender debiasing, reliance on rule-based or manual methods poses limitations in capturing nuanced biases and scalability. These approaches may compromise contextual meaning, leading to information loss and distorted communication. Motivated by these challenges, we leverage automated models to generate gender-neutralized texts, offering more effective and scalable solutions with potential for widespread impact.

We define *Gender-neutralized revision (GNR)* in the context of this research as the process of modifying a gender-biased text to remove gender-specific language or stereotypes while retaining the overall context and meaning. This involves replacing gendered terms with gender-neutral alternatives to promote fair and inclusive communication.

Table 1 presents examples from our *GENRE* corpus, where GB indicates gender-biased texts, and GNR denotes gender-neutralized revisions. In the first example, the revision removes the focus on gender and promotes equal treatment for all individuals. In the second example, the revised text eliminates gender bias and acknowledges that parking difficulties can be experienced by anyone, regardless of gender. The revision in the third example avoids singling out a specific gender and focuses on addressing the behavior or atti-

*Work was carried out at IIT Patna while doing internship
†The authors contributed equally and are the joint first authors.

tude of individuals rather than generalizing it to a particular group. In the fourth example, the revision removes the gender-specific term and acknowledges that humor can vary among individuals regardless of their gender. In the last case, the revised text avoids making generalizations based on gender and recognizes that sexual desires can differ among individuals. These gender-neutral revisions create inclusive communication, preventing stereotype reinforcement and reducing marginalization. They foster respectful dialogue, promoting understanding while avoiding gender-based biases.

The primary contributions are outlined as follows:

1. Catalyzing positive social transformation via inclusive and equitable text through gender-bias mitigation.
2. Introducing the unique *GEnder Neutralized REvisions (GENRE)* corpus from gender-stereotyped texts.
3. Proposing the *GEnder-NEutralizing Text Transformation (GENETT)* framework, leveraging encoders, vector quantization, and Neutrality-Infused Stylization (NIS) module for gender-neutral content generation.
4. Validating the model's effectiveness through extensive experiments on the *GENRE* corpus.

Task Alignment with UN SDGs and LNOB. This research aligns with several United Nations Sustainable Development Goals (UN SDGs), contributing to global sustainability and equality. By addressing gender bias in textual data and promoting gender-neutral content generation, the study supports *'Goal 5: Gender Equality'*, fostering inclusivity in language and society. Additionally, by mitigating gender bias and promoting inclusive communication, the research contributes to *'Goal 10: Reduced Inequalities'*, helping to reduce societal disparities and promote fairness. Moreover, by advocating for gender-neutral communication and respectful dialogue, the study supports *'Goal 16: Peace, Justice, and Strong Institutions'*, contributing to building peaceful and inclusive societies. Through its efforts to mitigate biases and promote inclusivity, this research aligns with the Leave No One Behind (LNOB) principle, ensuring that all individuals, regardless of gender identity or background, are considered and included in the journey towards sustainable development.

2 Related Work

Research on gender bias spans various domains, including artificial intelligence, language usage, job advertisements, and machine translation. This section reviews past research, emphasizing contributions, recognizing limitations, and showcasing our work on automated gender-neutral text generation as a response to these gaps. Previous works on coreference resolution, exemplified by [Zhao *et al.*, 2018a], introduced the WinoBias dataset, revealing bias in gendered pronoun associations. The authors proposed data augmentation and word-embedding debiasing techniques to rectify bias in the WinoBias corpus. In extending this analysis to languages with grammatical gender, [Zhou *et al.*, 2019] introduced novel bias definitions and leveraged bilingual word embeddings for analysis and mitigation. Additionally, [Zhao *et al.*, 2019] examined gender bias in ELMo's contextualized word embedding, suggesting methods to mitigate the observed bias.

[Yang and Feng, 2020] addressed bias in word embedding relations, proposing a causal approach to reduce gender bias by considering the statistical dependency between gender-definition and gender-biased word embedding. [Chiril *et al.*, 2020] focused on automatically detecting and characterizing sexist content on social media, utilizing speech acts theory and discourse analysis. For language models, [Garimella *et al.*, 2021] explored the mitigation of social biases, including gender bias, in BERT models by addressing biases in both model representations and generated text.

Gender biases in job advertisements were exposed by [Hu *et al.*, 2022], showing how implicit gender traits can perpetuate preferences and hinder gender equality. In machine learning, [Badaloni and Rodà, 2022] addressed the underrepresentation of female students in STEM, especially computer science. [Cohen *et al.*, 2023] observed improved performance for women in quantitative questions with gender-neutral language, highlighting its benefits without delving into text generation. [Piergentili *et al.*, 2023] discussed gender inclusivity in machine translation but lacked a concrete approach for automated gender-neutral text generation.

While these studies make significant contributions to the field, they do not directly address the automated generation of gender-neutral text. In contrast, our work focuses specifically on developing an automated model for generating gender-neutral versions of gender-biased text using advanced natural language processing and deep learning techniques. This approach fills a gap in existing literature, providing a practical solution for transforming biased text into inclusive, unbiased forms. By addressing limitations in previous methods, our approach fosters fairness and diminishes biases in natural language processing applications.

3 Dataset

We introduce the first-of-its-kind, corpus of *GEnder Neutralized REvisions (GENRE)* created from their gender-stereotyped versions.

3.1 Data Collection

GENRE is created by consolidating gender-biased instances (only English) from three benchmark datasets:

1. Workplace Sexism [Grosz and Conde-Cespedes, 2020], which contains 1100+ examples of workplace sexism, filtering out rare scenarios, removing duplicates, and using formal language;
2. Call Me Sexist [Samory *et al.*, 2021], retrieved from Twitter's Search API using the phrase "call me sexist, but", and annotated through crowd-sourcing;
3. *EXIST@IberLEF* [Rodríguez-Sánchez *et al.*, 2021], which compiled prevalent sexist terms and phrases in English and Spanish extracted from Twitter messages commonly undervaluing women's roles in society.

Table 2 compares various sexist datasets. None of the existing datasets is marked with neutralized revisions of the gender-biased instances, and *GENRE* corpus is the first of its kind. The corpus comprises of all instances (5230 instances) of the sexist class from the "Workplace Sexism" (627

instances), "EXIST 2021" (2794 instances), and the "Call Me Sexist" (1809 instances) datasets.

Datasets	Bias Labels	Size	GNR
Waseem & Hovy [Waseem and Hovy, 2016]	Racist, Sexist	3383	x
AMI@IberEval [Fersini *et al.*, 2018]	Misogynous	1851	x
Exist@IberLEF	Sexist	2794	x
Call me Sexist	Sexist	1809	x
GENRE (Ours)	**Gender-biased Stereotypes (GS)**	**5230**	✓

Table 2: Comparisons of different sexist datasets

3.2 Data Annotation

The data annotation process for gender-neutralizing the given gender-biased sentences involves three annotators, two linguists with Ph.Ds, and one computer science with Ph.D. Before starting the annotation process, the annotators were provided with clear guidelines and instructions on the task at hand. They were trained on the concepts of gender bias, the importance of gender-neutral language, and the goal of creating gender-neutralized versions of sentences. Each annotator was assigned a mutually exclusive set of gender-biased sentences for gender-neutralization, ensuring independent contributions. The aim is to remove gender-specific language and make the sentences more inclusive and unbiased.

Gender-Neutralization Process

- *Identifying Gender-Specific Terms:* The annotators read each sentence carefully to identify gender-specific terms like "men," "women", "he", "she", etc.
- *Replacing Gender-Specific Terms:* The annotators replace gender-specific terms with gender-neutral alternatives. For example, "men" may be replaced with "everyone", "he" or "she" with "they", etc.
- *Restructuring Sentences:* In some cases, the sentence structure may need to be altered to remove gender references. For instance, "Please women. When you commit to your man, commit to the gym as well. Commit to the kitchen too." is transformed into "Please, when you commit to someone, commit to self-improvement as well. Commit to self-sufficiency too".
- *Ensuring Sentence Coherency:* The annotators ensure that the revised sentences are grammatically correct and maintain the intended meaning of the original sentences.

Conflict Resolution: In case of any discrepancies or differences in gender-neutralization among the annotators, a senior annotator reviewed the conflicting sentences and decided on the most appropriate gender-neutralized version.

Review and Consistency Check.

Our approach involves independent gender-neutralization by each annotator, followed by a rigorous assessment of inter-annotator agreement using well-defined metrics. We conducted a thorough analysis of the GNRs using Fluency (F) and Adequacy (A) metrics. Annotator A1's gender-neutralizations were assessed by both A2 and A3 on a scale of 1-5 (Fluency - 5: Flawless, 4: Good, 3: Non-native, 2: Disfluent, 1: Incomprehensible; Adequacy - 5: All, 4: Most, 3: Much, 2: Little, 1: None). Likewise, A2's revisions were evaluated by A1 and A3, and the same process applied to A3. The scores reported in the Table 3 for each annotator reflect the average ratings given by the other two annotators during the review process. The high fluency and adequacy scores validate the quality and consistency of the generated GNRs.

HE	A1	A2	A3	Average
Fluency (F)	4.05	4.18	4.22	4.15
Adequacy (A)	3.95	4.11	3.83	3.96

Table 3: Quality scores from human evaluation of the GNRs.

Annotation Samples.

Here are some sample instances from the *GENRE* corpus:

1. **Actual Gender-biased Sentence:** *Just dress sexy and you will sign any contract.*
   Annotated GNR: *Just dress confidently and you will sign any contract.*
2. **Actual Gender-biased Sentence:** *I'm sorry but girls are just really unfunny.*
   Annotated GNR: *I'm sorry, but some individuals are just not funny.*
3. **Actual Gender-biased Sentence:** *She is a soured woman, she needs a man in her life!*
   Annotated GNR: *She is unhappy, she may need support in her life!*

4 Methodology

We design a *GEnder-NEutralizing Text Transformation (GENETT)* framework for the generation of gender-neutralized versions of gender-biased texts.

4.1 Task Definition

The problem entails training a model to generate gender-neutral versions of biased text. This involves transforming the biased text into a form that avoids reinforcing gender stereotypes or biases while maintaining the original context and meaning. Mathematically, this can be defined as:

Let b be the input gender-biased text, and n be the corresponding gender-neutral output text generated by the automated model. The objective is to find a function $\mathcal{F}$ that transforms b to n: $n = \mathcal{F}(b)$

The model's training process involves learning the relationships between gender-biased and gender-neutral expressions. This involves capturing syntactic and semantic features in the text and understanding the context in which gender-specific terms are used. During inference, the trained model takes a gender-biased text as input and generates a gender-neutralized version.

4.2 Proposed Framework

Figure 1 illustrates the general architecture of our proposed approach. The *GENETT* framework consists of four encoders for extracting continuous and quantized features of biased and neutral texts, two codebooks to store cluster centers representing the distributions of biased and neutral texts, and

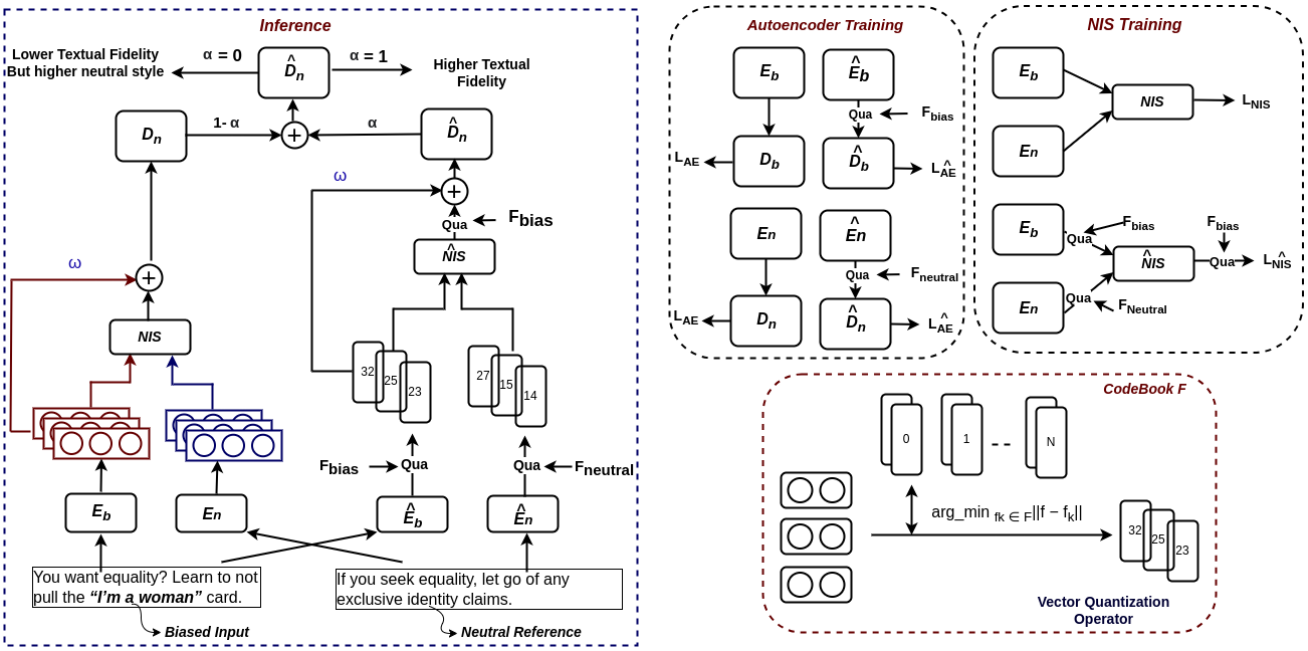

Figure 1: Illustration of the *GEnder-NEutralizing Text Transformation (GENETT)* framework.

two Neutrality-Infused Stylization (NIS) modules for effective style transfer.

Data Collection and Pre-processing. We begin by extracting the features of the biased text, denoted as *b*, and the neutral text, denoted as *n*, using two BART [Lewis *et al.*, 2019] encoders labeled as E_b and E_n. These encoders are responsible for capturing and encoding the essential information from the input texts. Subsequently, we decode these extracted features back into their respective texts b_{re} and n_{re}. Specifically, the features of the biased text are decoded using a BART decoder D_b, while the features of the neutral text are decoded using another BART decoder D_n:

$$b_{re} = D_b(E_b(b)), \quad n_{re} = D_n(E_n(n)) \tag{1}$$

To optimize the encoder and decoder, the reconstruction loss ($\mathbf{L}_{recon}$) is defined as follows:

$$\mathbf{L}_{recon} = ||b_{re} - b|| + \mathbf{L}_a(E_b, E_b, \mathcal{D}_b) \tag{2}$$

where $\mathbf{L}_a$ is the adversarial loss and $\mathcal{D}_b$ is the discriminator network:

$$\mathbf{L}_a(E_b, E_b, \mathcal{D}_b) = ||\log \mathcal{D}_b(b) + \log(1 - \mathcal{D}_b(b_{re}))|| \tag{3}$$

The optimization of the encoder and decoder for the neutral text, E_n and D_n, is performed in the same manner as for E_b and D_b.

Quantization and Codebooks. To achieve compressed and discrete representation, we utilize vector quantization and build two codebooks denoted as $F_{bias} \in \mathbf{R}^{N \times D}$ and $F_{neutral} \in \mathbf{R}^{N \times D}$, where N represents the number of entries, and D denotes the dimension of each entry. To enhance the representation performance of the quantized features, we incorporate two additional encoders, denoted as $\hat{E}_b$ and $\hat{E}_n$, which extract the features of the quantized features:

$$b_f = \hat{E}_b(b) \quad \text{and} \quad n_f = \hat{E}_n(n) \tag{4}$$

To obtain the quantized features $\hat{b}_f$ and $\hat{n}_f$, we perform vector quantization using the codebooks F_{bias} and $F_{neutral}$:

$$\hat{b}_f = \mathbf{Q}_{F_{bias}}(b_f), \quad \hat{n}_f = \mathbf{Q}_{F_{neutral}}(n_f) \tag{5}$$

where $\mathbf{Q}_F(f) = \arg\min f_k \in F||f - f_k||$ is the vector quantization operator that substitutes the original features with the nearest entry from the codebook F. The quantized features are then decoded into textual features $\hat{b}_{re}$ and $\hat{n}_{re}$ using decoders $\hat{D}_b$ and $\hat{D}_n$:

$$\hat{b}_{re} = \hat{D}_b(\hat{b}_f) \quad \text{and} \quad \hat{n}_{re} = \hat{D}_n(\hat{n}_f) \tag{6}$$

To optimize the codebooks F_{bias} and $F_{neutral}$ jointly with the reconstruction loss, we follow a similar approach as described in Equation (5). The quantized reconstruction loss is defined as:

$$\begin{aligned}\hat{\mathbf{L}}_{recon}(\hat{E}_b, \hat{D}_b, F_{bias}) = \hat{\mathbf{L}}_{recon}(\hat{E}_b, \hat{D}_b) + \\ ||\text{gr}[\hat{b}_f] - b_f|| + ||\text{gr}[b_f] - \hat{b}_f||\end{aligned} \tag{7}$$

where gr is the stopping gradient. The optimization of the codebook F_{bias} is achieved through the second term, which refines and improves the codebook representation during training. The third term ensures alignment of the latent feature b_f with the nearest neighbor entry in the codebook, promoting alignment with existing representations.

The optimization of the variables $\hat{E}_n$, $\hat{D}_n$, and $F_{neutral}$ is performed using the same loss function $\hat{\mathbf{L}}_{recon}(\hat{E}_n, \hat{D}_n, F_{neutral})$. This ensures a consistent and unified optimization approach for all the involved components in the framework.

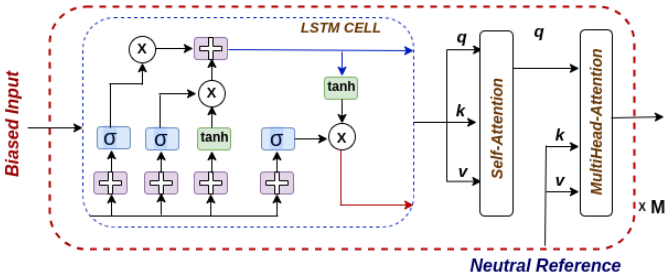

Figure 2: Illustration of the NIS module

Neutrality-Infused Stylization (NIS). We introduce the Neutrality-Infused Stylization (NIS) module to ensure effective style transfer for both continuous and quantized features. This module, comprising an LSTM cell and two attention blocks with residual connections, takes biased text feature $B_f \in \mathbf{R^d}$ and neutral text feature $N_f \in \mathbf{R^d}$ as inputs, generating the stylized feature vector $Y_f \in \mathbf{R^d}$. (See Figure 2 for an illustration of the NIS module's functionality.)

The attention blocks (AT) operate on query (q), key (k), and value (v) inputs:

$$AT(q, k, v) = \text{Softmax}(E_q(q) \cdot E_k(k)) \cdot E_v(v) + q \quad (8)$$

E_q, E_k, and E_v denote embedding layers. The first attention block utilizes self-attention with input heads q, k, and v, which receive the transformed content feature $\hat{B_f} = \text{LSTM}(B_f)$ from an LSTM layer applied to the biased text feature B_f. In the second attention block, $\hat{B_f}$ serves as q, while N_f acts as k and v, allowing the module to attend to both neutral information in N_f and contextual information in $\hat{B_f}$. The stylized feature Y_f is computed accordingly.

$$Y_f = \text{NIS}(B_f, N_f) = AT(AT(\hat{B_f}, \hat{B_f}, \hat{B_f}), N_f, N_f) \quad (9)$$

The NIS module includes an additional self-attention block to capture global information from the quantized codes, facilitating a thorough analysis of the entire code sequence for long-range dependencies and contextual information. The attention blocks also feature residual connection, which helps maintain the integrity of content details during the style transfer process. By preserving the connection between the input and output, important information is retained, ensuring the generated output remains faithful to the original content.

To enhance style transfer performance, we iteratively apply the NIS module multiple times, refining the process for improved quality and accuracy. The NIS module is trained with bias loss $\mathbf{L}_b$, adversarial loss $\mathbf{L}_a$, and neutral loss $\mathbf{L}_n$.

$$\mathbf{L}_{NIS} = \mathbf{L}_b + \mathbf{L}_a + \mathbf{L}_n \quad (10)$$

where $\mathbf{L}_b$ measures the difference between the stylized feature Y_f and the neutral feature N_f, $\mathbf{L}_a$ is an adversarial loss to encourage alignment with the distribution of style reference features, and $\mathbf{L}_n$ encourages alignment of mean and standard deviation with the neutral features.

For the transfer of quantized features, we employ an additional NIS module:

$$\hat{Y_f} = \hat{NIS}(\hat{B_f}, \hat{N_f}) \quad (11)$$

where $NIS(.,.)$ and $\hat{NIS}(.,.)$ have similar architectures but different parameters. The $\hat{NIS}$ module is optimized with a quantized reconstruction loss, given by $\mathbf{L}_{\hat{NIS}} = \mathbf{L}_{NIS} + ||\text{gr}[\hat{Y_f}] - Y_f||$.

Inference. In the inference step of the GENETT framework, we extract continuous and quantized features from biased and neutral texts using their respective encoders. These features undergo transformation via the NIS module, producing stylized continuous feature Y_f and stylized quantized feature $\hat{Y}_f$, preserving the reference style while maintaining content fidelity. The equilibrium among bias, neutral text, and textual fidelity is realized by the following equation.

$$\mathbf{z}_{test} = \mathcal{W}_\alpha(\mathcal{W}_\omega(\hat{Y_f}, \hat{b}_f), \mathcal{W}_\omega(Y_f, b_f)) \quad (12)$$

Here, $\mathcal{W}_l(c, d) = l \cdot c + (1 - l) \cdot d$, where ω controls fidelity (higher α means higher textual fidelity). Customizing the fidelity trade-off is flexible based on factors, such as input text and user preferences, facilitated by the *GENETT* framework, denoted as *GENETT*(α, β). Users can customize the balance between fidelity to the original text and desired style transfer by adjusting parameters α and β in *GENETT*. This flexibility enables a personalized trade-off, ensuring users achieve the desired balance between fidelity and style transfer in their experience with *GENETT*(α, β).

Component Design Justification. The framework components are designed to tackle the multifaceted nature of gender bias in text, aiming to identify, transform, and neutralize biased language while preserving context.

- This involves using four encoders to extract both continuous and quantized features from biased and neutral texts separately, enabling the capture of distinct characteristics for each text type.
- Quantization reduces computational complexity and memory footprint by achieving compressed representations, aided by two codebooks storing distribution-specific cluster centers for biased and neutral texts.
- Additionally, the neutrality-infused stylization module facilitates style transfer while ensuring gender neutrality, mitigating bias in language elements like word choice and tone through two NIS modules that attend to both biased and neutral features.

5 Experiments

We evaluate our model's performance on the GENRE dataset drawing comparisons with ten state-of-the-art systems on the same, namely: DRLST [John *et al.*, 2018], BST [Prabhumoye *et al.*, 2018], CAAE [Shen *et al.*, 2017], Ctrl-Gen [Hu *et al.*, 2017], ARAE [Zhao *et al.*, 2018b], Multi-Dec [Fu *et al.*, 2018], Style-Emb [Fu *et al.*, 2018], Point-Then-Operate (PTO) [Wu *et al.*, 2019], DualRL [Luo *et al.*, 2019] and PFST [He *et al.*, 2021]. To measure content preservation, we adopt BLEU, ROUGE, and, Cosine Similarity (CS). We compute the perplexity score (PPL) to quantify the fluency of the transferred sentences for finding BLEU and sBLEU we use *multi-bleu.perl*[1].

[1]https://github.com/OpenNMT/OpenNMT/blob/master/benchmark/3rdParty/multi-bleu.perl

6 Results and Analysis

Table 4 presents the performance of the *GENETT* framework on the *GENRE* dataset.

6.1 Implementation Details

We employ PyTorch[2], a Python-based deep learning framework, to construct our proposed model. Our experiments utilize BART, imported from the huggingface transformers[3] package. The codebook comprises N = 1024 entries, each with a dimension of d = 256. Our style transfer model integrates six NIS modules. All experimentation takes place on an NVIDIA GeForce RTX 2080 Ti GPU. We perform a grid search spanning 200 epochs and apply k-cross-validation[4].

For optimization, we employ the Adam algorithm [Kingma and Ba, 2015] with a learning rate of 0.05 and a dropout rate of 0.5. The LSTM cell dimension is determined as 812 through empirical analysis. The discriminator $\mathcal{D}$ encompasses two fully connected layers along with a ReLU layer, processing 812-dimensional input features. Stochastic gradient descent operates with a learning rate of 2e-4, a weight decay of 1e-3, and a momentum of 0.5. The average training time of the proposed GENETT framework on a NVIDIA GeForce RTX 2080 Ti GPU (11GB GDDR6) is around 157 minutes, which is roughly around 188.4 seconds/epoch.

6.2 Comparison with Existing Works

In comparison to baseline models, *GENETT* demonstrates notable enhancements in content preservation (BLEU) and fluency (PPL). While DualRL and PFST exhibit competitive content preservation (BLEU scores), they suffer a more noticeable drop in fluency (higher PPL scores). Style-Emb achieves enhanced fluency (low PPL scores) but experiences a significant decrease in content preservation (lower BLEU scores), indicating a trade-off. PTO's content preservation matches *GENETT* (high BLEU scores), but its fluency improvement (low PPL scores) is moderate.

GENETT's consistent and substantial improvement across all metrics results in contextually accurate (content preservation) and linguistically coherent (fluency) text, reaffirming its effectiveness in addressing gender bias and performing sentiment style transfer.

6.3 Human Evaluation

To assess the quality of gender-neutral texts generated by *GENETT*, a human evaluation was conducted on 300 randomly chosen instances from the test set of a random cross-validation fold. Four well-defined metrics were used for the assessment process [Singh *et al.*, 2022], and a score ranging from 0 to 5 was awarded based on these metrics. The most incorrect responses received a score of 0, while the best received a score of 5. The evaluator examined Fluency, Adequacy, Knowledge Consistency (KC), and Informativeness (Inf). PTO and Style-Emb exhibit moderate consistency across all metrics. PFST faces challenges in fluency, adequacy, and knowledge consistency. DualRL excels in fluency and knowledge consistency but lags in adequacy and informativeness. *GENETT* outperforms other models, receiving the highest scores in all four metrics, indicating its strength in generating responses that are fluent, semantically adequate, consistent, and highly informative.

Models	BLEU	ROUGE	PPL	CS
DRLST	38.7	0.39	1.78	0.84
BST	35.6	0.39	2.32	0.83
CAAE	31.4	0.34	1.94	0.86
Ctrl-Gen	36.1	0.40	2.11	0.85
ARAE	36.44	0.35	1.64	0.88
Multi-Dec	37.60	0.33	1.70	0.78
Style-Emb	39.4	0.39	2.68	0.83
PTO	46.8	0.46	2.67	0.89
DualRL	48.5	0.44	1.65	0.91
PFST	46.6	0.42	1.67	0.90
GENETT	**54.1**	**0.50**	**1.24**	**0.93**

Table 4: Performance comparison of the various models on the *GENRE* dataset. Bold values represent the maximum scores.

Models	Fluency	Adequacy	KC	Inf
PTO	3.08	3.11	3.04	2.87
Style-Emb	3.16	3.05	3.14	2.86
PFST	2.91	2.75	2.85	2.85
DualRL	3.21	2.89	3.25	2.81
GENETT *(ours)*	**3.36**	**3.41**	**3.32**	**3.83**

Table 5: Results of human evaluation. Here, KC: Knowledge Consistency, Inf: Informativeness, F: Fluency

6.4 Ablation Study

Ablation experiments (Table 6) involve removing specific components. In order to emphasize the significance of quantization, we conduct an experiment (*GENETT - [Quantizations]*) by eliminating all quantization vectors. In this case, we solely employ the end-to-end BART model [Lewis *et al.*, 2019] with the NIS module excluded. Quantizations impact fluency positively, but removal increases perplexity, indicating a trade-off. Next, to underscore the relevance of the NIS module, we perform a comparison by removing the module and directly concatenating the output from both encoders. The complete NIS module crucially preserves content across metrics, with its removal causing substantial drops in quality. This combined output is then passed to the decoders (*GENETT - [NIS]*). Following that, whe LSTM cell within the NIS module is removed (*GENETT - [LSTM in NIS]*), and instead, we directly pass the embeddings through the self-attention layer along with the reference input. To emphasize the role of self-attention within the NIS module (*GENETT - [Self-Attention in NIS]*), we contrast this approach with one where we concatenate the text embedding with the reference input. Self-attention in NIS has a minor impact, while LSTM in NIS significantly influences both content preservation and fluency. Overall, a trade-off between these two factors is evident. Quantization, self-attention, LSTM, and NIS all positively contribute to content preservation metrics, and their removal decreases content preservation quality.

[2]https://pytorch.org/

[3]https://huggingface.co/docs/transformers/index

[4]In our case, k is empirically set to 5

Setup	*GENRE (ours)*			
	BLEU	***ROUGE***	***CS***	***PPL***
GENETT - [Quantizations]	51.43 (-2.67)	0.47 (-0.03)	0.90 (+0.03)	1.39 (+0.15)
GENETT - [Self-Attention in NIS]	52.77 (-1.33)	0.49 (-0.01)	0.91 (+0.02)	1.25 (+0.01)
GENETT - [LSTM in NIS]	52.13 (-1.97)	0.48 (-0.02)	0.91 (+0.02)	1.28 (+0.04)
GENETT - [NIS]	51.93 (-2.17)	0.46 (-0.04)	0.89 (+0.04)	1.24 (0)

Table 6: Results illustrating variance in outputs of the *GENETT* model when individual components are ablated.

1. Original Sentence:	**"I'm going to miss her resting bitch face".**
GNR [HA]	**I will miss her unpleasant expression.**
GNR [Multi-Dec]	I will miss this expression.
GNR [Style-Emb]	I'm going to miss individual resting face.
ChatGPT	I'm going to miss their neutral expression.
GNR [GENETT] *(ours)*	I will miss her unpleasant expression.
2. Original Sentence:	**"What a dumb bitch".**
GNR [HA]	**What an unintelligent individual!**
GNR [Multi-Dec]	what a person
GNR [Style-Emb]	What a dumb
ChatGPT	What a dumb individual.
GNR [GENETT] *(ours)*	What an unintelligent individual!

Table 7: Sample predictions from different systems. HA: Human Annotator.

6.5 Qualitative Analysis

Table 7 compares GNRs from the best-performing baseline models (Multi-Dec and Style-Emb) with our proposed GENETT model and the human annotator's gold standard. In example 1, Multi-Dec falls short compared to the human annotator's version, while Style-Emb introduces unnatural phrasing and loses sentiment alignment. In contrast, *GENETT* successfully captures gender-neutralization while preserving both the intended meaning and sentiment, aligning closely with the gold standard. In example 2, Multi-Dec, although accurate in gender-neutralization, fails to fully capture the tone and intensity of the original sentiment. The Style-Emb GNR loses the original sentiment due to truncation. Conversely, *GENETT* skillfully retains the sentiment while effectively neutralizing gender bias. Limitations observed in the baselines highlight challenges in balancing sentiment preservation, fluency, and gender neutralization. *GENETT* excels in addressing these challenges, consistently producing GNRs that are linguistically accurate, coherent, and contextually faithful to the original intent.

Comparison with ChatGPT-3.5

We recognize the significance of conducting a comparative analysis with LLMs (such as ChatGPT 3.5[5]) for gender neutralization, as such an analysis can provide a deeper understanding of the effectiveness of our approach. After meticulous prompt optimization, we utilized the following prompt for one-shot training to generate gender-neutralizing texts with ChatGPT. Here is the prompt generating two GNRs:

Prompt: We define Gender-neutralized revision (GNR) in the context of this research as the process of modifying a gender-biased text to remove gender-specific language or stereotypes while retaining the overall context and meaning. This involves replacing gendered terms with gender-neutral alternatives to promote fair and inclusive communication. For example, the sentence "Chicks just aren't that funny." can be written as "Some people just aren't that funny." after gender neutralizing revision.

Based on the above description of GNR, generate gender-neutralized version of the following sentences:

1. I'm going to miss her resting bitch face.
2. What a dumb bitch.

ChatGPT Output:

1. I'm going to miss their neutral expression.
2. What a dumb individual.

After evaluating ChatGPT's performance on a sample of 500 instances from our dataset, we observed its effective transformation of gender-specific sentences into gender-neutral ones. However, there are occasions when it alters the intended meaning of the generated output, which is not our intended objective. As evident in the first example, ChatGPT erroneously replaced 'her' with 'their', resulting in a change of the target entity and a subsequent alteration of the sentence's meaning. This underscores the necessity of manual annotations and diligent supervision to effectively accomplish the task of eliminating gender bias while accurately preserving the intended meaning.

7 Conclusion

This study has made significant strides in advancing the field of gender-neutral text generation through the introduction of the *GENETT* framework. By leveraging auto-encoders, vector quantization, and the Neutrality-Infused Stylization (NIS) module, *GENETT* offers a comprehensive solution for addressing gender bias in textual data. Extensive experiments demonstrate the framework's effectiveness, showing significant improvements in content preservation and fluency over state-of-the-art methods. Our findings indicate that *GENETT* successfully neutralizes gender bias outperforming existing models and even large language models like ChatGPT-3.5 in various metrics. The *GENRE* corpus serves as a valuable resource for further research and applications in gender-bias analysis and mitigation.

Future research should focus on extending the framework to various domains and languages to promote broader inclusivity and unbiased communication. Additionally, integrating gender-neutral language practices into broader societal contexts and raising awareness about the importance of inclusive language use will further support efforts towards reducing societal disparities and promoting fairness.

[5]https://chat.openai.com/

Ethical Statement

This section addresses the ethical aspects of our research, focusing on data collection practices, cultural sensitivity, and the limitations of gender representation. It highlights the importance of transparency, inclusivity, and adherence to ethical standards in mitigating gender bias in textual data.

Data Collection and Availability

This study utilized three openly accessible benchmark datasets to construct the GENRE corpus, ensuring strict adherence to copyright regulations. Access to the code and data is provided for research purposes through a suitable data agreement mechanism.

Cultural and Contextual Sensitivity

Gender bias varies significantly across different cultures and contexts. Applying gender-neutralization techniques without considering these variations may result in inappropriate or insensitive translations. While the goal is to promote inclusivity, modifying texts to be gender-neutral could potentially raise concerns about restricting free speech or altering the original intentions of authors. It is crucial to balance bias reduction with preserving the authenticity of the author's voice. Users should be informed when interacting with content modified by an AI system. Transparency about such interventions is vital to maintaining trust and accountability in communication.

Limitations

Scope and Gender Representation

The primary focus of this work is to enhance the understanding of cyber harassment conversations, particularly in underresearched areas. Understanding these dialogues is a crucial step towards addressing and mitigating cyber harassment, which affects individuals irrespective of their gender identity. The exclusion of non-binary cases in this study is due to the lack of suitable datasets, not insensitivity. Our paper employs binary gender pronouns such as "He" and "She" for data annotation, aligning with the existing datasets that predominantly use binary gender classifications. Recognizing the limitations of these datasets is essential, and we acknowledge the importance of non-binary individuals and they/them pronouns. Future research aims to include these elements when appropriate datasets become available.

References

[Badaloni and Rodà, 2022] Silvana Badaloni and Antonio Rodà. Gender knowledge and artificial intelligence (short paper). In Guido Boella, Fabio Aurelio D'Asaro, Abeer Dyoub, and Giuseppe Primiero, editors, *Proceedings of 1st Workshop on Bias, Ethical AI, Explainability and the Role of Logic and Logic Programming (BEWARE 2022) co-located with the 21th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2022), Udine, Italy, December 2, 2022*, volume 3319 of *CEUR Workshop Proceedings*, pages 107–112. CEUR-WS.org, 2022.

[Chiril *et al.*, 2020] Patricia Chiril, Véronique Moriceau, Farah Benamara, Alda Mari, Gloria Origgi, and Marlène Coulomb-Gully. He said "who's gonna take care of your children when you are at acl?": Reported sexist acts are not sexist. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4055–4066. Association for Computational Linguistics, 2020.

[Cohen *et al.*, 2023] Alma Cohen, Tzur Karelitz, Tamar Kricheli-Katz, Sephi Pumpian, and Tali Regev. Gender-neutral language and gender disparities. *NBER Working Paper*, (w31400), 2023.

[Fersini *et al.*, 2018] Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. Overview of the task on automatic misogyny identification at ibereval 2018. *Ibereval@ sepln*, 2150:214–228, 2018.

[Fu *et al.*, 2018] Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. Style transfer in text: Exploration and evaluation. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 663–670. AAAI Press, 2018.

[Garimella *et al.*, 2021] Aparna Garimella, Akhash Amarnath, Kiran Kumar, Akash Pramod Yalla, Anandhavelu Natarajan, Niyati Chhaya, and Balaji Vasan Srinivasan. He is very intelligent, she is very beautiful? on mitigating social biases in language modelling and generation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 4534–4545. Association for Computational Linguistics, 2021.

[Grosz and Conde-Cespedes, 2020] Dylan Grosz and Patricia Conde-Cespedes. Automatic detection of sexist statements commonly used at the workplace. In *Trends and Applications in Knowledge Discovery and Data Mining: PAKDD 2020 Workshops, DSFN, GII, BDM, LDRC and LBD, Singapore, May 11–14, 2020, Revised Selected Papers 24*, pages 104–115. Springer, 2020.

[He *et al.*, 2021] Junxian He, Xinyi Wang, Graham Neubig, and Taylor Berg-Kirkpatrick. A probabilistic formulation of unsupervised text style transfer. *arXiv preprint arXiv:2002.03912*, 2021.

[Hu *et al.*, 2017] Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. Toward controlled generation of text. In *International conference on machine learning*, pages 1587–1596. PMLR, 2017.

[Hu *et al.*, 2022] Shenggang Hu, Jabir Alshehabi Al-Ani, Karen D. Hughes, Nicole Denier, Alla Konnikov, Lei Ding, Jinhan Xie, Yang Hu, Monideepa Tarafdar, Bei

Jiang, Linglong Kong, and Hongsheng Dai. Balancing gender bias in job advertisements with text-level bias mitigation. *Frontiers Big Data*, 5:805713, 2022.

[John *et al.*, 2018] Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. Disentangled representation learning for non-parallel text style transfer. *arXiv preprint arXiv:1808.04339*, 2018.

[Kingma and Ba, 2015] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[Lewis *et al.*, 2019] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

[Luo *et al.*, 2019] Fuli Luo, Peng Li, Jie Zhou, Pengcheng Yang, Baobao Chang, Zhifang Sui, and Xu Sun. A dual reinforcement learning framework for unsupervised text style transfer. *arXiv preprint arXiv:1905.10060*, 2019.

[Piergentili *et al.*, 2023] Andrea Piergentili, Dennis Fucci, Beatrice Savoldi, Luisa Bentivogli, and Matteo Negri. From inclusive language to gender-neutral machine translation. *arXiv preprint arXiv:2301.10075*, 2023.

[Prabhumoye *et al.*, 2018] Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. Style transfer through back-translation. *arXiv preprint arXiv:1804.09000*, 2018.

[Rodríguez-Sánchez *et al.*, 2021] Francisco Rodríguez-Sánchez, Jorge Carrillo-de Albornoz, Laura Plaza, Julio Gonzalo, Paolo Rosso, Miriam Comet, and Trinidad Donoso. Overview of exist 2021: sexism identification in social networks. *Procesamiento del Lenguaje Natural*, 67:195–207, 2021.

[Samory *et al.*, 2021] Mattia Samory, Indira Sen, Julian Kohne, Fabian Flöck, and Claudia Wagner. "call me sexist, but...": Revisiting sexism detection using psychological scales and adversarial samples. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 573–584, 2021.

[Shen *et al.*, 2017] Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. Style transfer from non-parallel text by cross-alignment. *Advances in neural information processing systems*, 30, 2017.

[Singh *et al.*, 2022] Gopendra Vikram Singh, Mauajama Firdaus, Shruti Mishra, Asif Ekbal, et al. Knowing what to say: Towards knowledge grounded code-mixed response generation for open-domain conversations. *Knowledge-Based Systems*, 249:108900, 2022.

[Singh *et al.*, 2023] Gopendra Singh, Soumitra Ghosh, and Asif Ekbal. Promoting gender equality through gender-biased language analysis in social media. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 6210–6218, 2023.

[Waseem and Hovy, 2016] Zeerak Waseem and Dirk Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93, 2016.

[Wu *et al.*, 2019] Chen Wu, Xuancheng Ren, Fuli Luo, and Xu Sun. A hierarchical reinforced sequence operation method for unsupervised text style transfer. *arXiv preprint arXiv:1906.01833*, 2019.

[Yang and Feng, 2020] Zekun Yang and Juan Feng. A causal inference method for reducing gender bias in word embedding relations. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9434–9441. AAAI Press, 2020.

[Zhao *et al.*, 2018a] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. In Marilyn A. Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 15–20. Association for Computational Linguistics, 2018.

[Zhao *et al.*, 2018b] Junbo Zhao, Yoon Kim, Kelly Zhang, Alexander Rush, and Yann LeCun. Adversarially regularized autoencoders. In *International conference on machine learning*, pages 5902–5911. PMLR, 2018.

[Zhao *et al.*, 2019] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. Gender bias in contextualized word embeddings. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 629–634. Association for Computational Linguistics, 2019.

[Zhou *et al.*, 2019] Pei Zhou, Weijia Shi, Jieyu Zhao, Kuan-Hao Huang, Muhao Chen, and Kai-Wei Chang. Analyzing and mitigating gender bias in languages with grammatical gender and bilingual word embeddings. *ACL: Montréal, QC, Canada*, 2019.