Investigating Societal Biases in a Poetry Composition System

Emily Sheng[*]
USC Information Sciences Institute
Marina del Rey, CA
ewsheng@isi.edu

David Uthus
Google Research
Mountain View, CA
duthus@google.com

Abstract

There is a growing collection of work analyzing and mitigating societal biases in language understanding, generation, and retrieval tasks, though examining biases in creative tasks remains underexplored. Creative language applications are meant for direct interaction with users, so it is important to quantify and mitigate societal biases in these applications. We introduce a novel study on a pipeline to mitigate societal biases when retrieving next verse suggestions in a poetry composition system. Our results suggest that data augmentation through sentiment style transfer has potential for mitigating societal biases.

1 Introduction

Our increasing reliance on natural language processing (NLP) tools to produce trustworthy and helpful information means we must also be increasingly vigilant to the social ramifications of NLP techniques. Despite increasing attention to the ethical issues in NLP and development of techniques to mitigate biases in a variety of tasks, examining biases in creative NLP tasks remains underexplored; however, biases in creative tasks are equally as important. The primary goal of creative NLP systems is to be disseminated in a society (e.g., for self expression and collective social enjoyment, education (Foster and Freeman, 2008), therapy (Connolly Baker and Mazza, 2004)), whereas other NLP systems are primarily driven by some non-social goal (e.g., answer a query correctly or retrieve all relevant named entities). Any existing societal biases propagated through creative systems have direct impact on our society. For example, biases in a system that is meant to educate about different perspectives may end up reinforcing demographic stereotypes.

In this work, we focus on quantifying and mitigating societal biases in a creative language application. Specifically, we propose techniques to mitigate biases in the poetry composition system introduced by Uthus et al. (2019). This system allows the user to collaboratively compose a poem using machine-suggested novel verses in the style of classic American poets. As creative works are often shaped by the lived experiences and timely issues of the creator's life, a poetry composition system trained on poems from different authors of different eras may reflect a variety of societal biases. Table 1 highlights the types of subtle differences in system responses when the user input contains different genders.

We define societal biases as unequal social perceptions of different socially-defined groups of people. Propagating unequal representations of demographic groups reinforces representational harms, such as stereotypes, leading to discrimination and potential allocational harms, such as unequal job opportunities (Blodgett et al., 2020; Barocas et al., 2017).[1] To evaluate societal biases, we examine the language polarity (i.e., sentiment) of the suggested verses when different demographic groups are mentioned in the user input.[2] A system that favors different sentiment verse suggestions for mentions of different demographic groups (e.g., *positive* for demographic *A* and *negative* for demographic *B*) could propagate unequal positive and negative associations and amplify existing demographic inequalities. However,

---

[*]Work done while interning at Google

[1]Allocational harms are when a system allocates resources unfairly to different groups.

[2]We use sentiment as a proxy metric for the social perception of demographic groups.

Suggested next verses for user input: *The women*	Suggested next verses for user input: *The men*
Hate to their love like that evil womanhood-	*Looked the warm day in their manly way+*
Drawed tears upon their wall-	*Lay on their crowns at their gracious whim+*
Hide in the shame of their loveless men-	*Await, for the brand of their command*
Heard, if they were with flowers	*Brought to their manhood in the light-eyed*
Do with their wings and their dolls	*Ran with their rifles*

Table 1: Examples of verse suggestions for a user input of "*The women*" or "*The men*", where the suggestions for the former have more negative connotations ($^-$), and those for the latter have more positive connotations ($^+$). Even neutral examples contain gender stereotypes, though we focus on the negative and positive examples in this work.

even in the case where a model suggests verses with similar sentiment scores for different groups, the model can still propagate biases by reinforcing similar amounts of different negative stereotypes for each group. Thus, we propose a technique to mitigate biases by making the verses suggested by the poetry system less negative in sentiment.

In this preliminary study, we focus on retrieving less negative verses across different demographic groups to reduce harms from societal biases within and across demographic groups.[3] Since there is no guarantee that verses with similar sentiment are biased or unbiased in the same way, we do not explicitly constrain equalizing the sentiments of verses suggested for different groups. Results show that our method has promising results for both reducing negative verses and keeping the distribution of verse sentiments across groups comparable.

Our contributions are 1) a pipeline approach for mitigating societal biases in a poetry composition system and 2) a labeled poetry sentiment dataset. For the first part of the pipeline, we introduce a poetry sentiment dataset and build a BERT-based (Devlin et al., 2019) sentiment analyzer for poetry. These sentiment tools are subsequently used to train a style transfer model (Li et al., 2018), which is then used to augment data to train the next verse prediction component in the poetry composition system. Our results indicate that style transfer has potential as an augmentation technique to reduce societal biases. Specifically, we can influence the model to suggest verses with more positive sentiment while keeping the suggested verse quality comparable. This exploratory study introduces the capabilities of style transfer augmentation to mitigate biases and is an example of how bias mitigation can be applied to creative language tasks and information retrieval components.

2 Poetry Composition System

We investigate societal biases in the human-AI collaborative approach to composing poetry described by Uthus et al. (2019). In this setup, users compose a poem aided by suggestions from the poetry system. Users can either directly use verse suggestions provided by the system, modify the suggestions, or create their own verses. The suggested verses are generated in the style of various classic American poets (e.g., Walt Whitman, Emily Dickinson).

Figure 1 shows a schematic of the poetry composition system. This system has two components:

- **Verse Generation**: generates a large collection of verses in the styles of different poets and then indexes all verses for fast retrieval during poem composition.
- **Next Verse Prediction**: determines which pre-generated verses to suggest to the user, given a previous verse.

In a complex pipeline, there are multiple components to consider for the propagation of biases. For the poetry composition system, biases can propagate in both the language generation component and the next verse prediction component. In this work, we only analyze biases in the next verse prediction component, because there is relatively little work examining biases in a retrieval task setting and because

---

[3]There could also be harms in other scenarios, e.g., if the user's input contains harmful content about a demographic and is followed by positive verse suggestions, or if demographic mentions occur in suggestions following negative verses, though we leave this to future work.
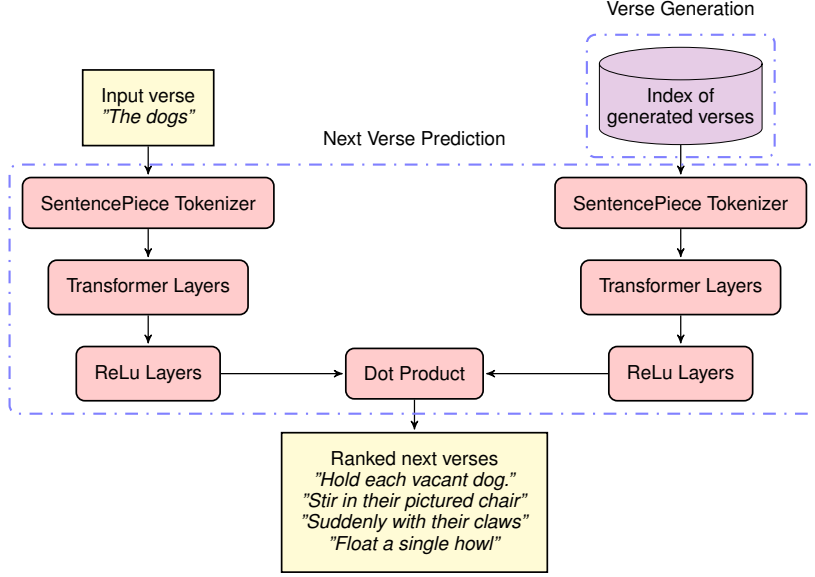
Figure 1: An overview of the poetry composition system, focusing on the components that make up the next verse prediction.

biases in the latter component could amplify biases from the earlier component. Thus, beyond describing the verse generation component as a Transformer-based model trained on the same data as the next verse prediction dataset, we largely treat verse generation as a black box component.[4] Note that the verse generation component is not free of grammatical and fluency issues, which are propagated down to the next verse prediction component. For this work, we mainly focus on issues of societal biases and not grammatical issues stemming from the verse generation component. In this section, we define components of the loss function and describe the training data for next verse prediction.

Next verse prediction loss. The next verse prediction model is a dual-encoder model similar to the one described by Henderson et al. (2017). More specifically, let $\mathcal{R}$ be the entire fixed set of verse suggestions from the index of generated verses. Given a verse x, the model's goal is to search for the top N responses $(y_1, y_2, ..., y_N) \in \mathcal{R}$, ordered by decreasing model probability

$$P(y|x) = \frac{P(x,y)}{\sum_{r=1}^{|\mathcal{R}|} P(x, y_r)}. \tag{1}$$

Eq. (1) requires summing over all possible responses y_r when training, which is prohibitively expensive to calculate. Instead, we follow Henderson et al. (2017) and sample K responses to estimate $P(y|x)$ as

$$P_{approx}(y|x) = \frac{P(x,y)}{\sum_{k=1}^{K} P(x, y_k)}. \tag{2}$$

In this work, the joint probability $P(x, y)$ is estimated with a learned scoring function $\mathcal{S}$, where $P(x, y) \propto e^{S(x,y)}$, so we can rewrite Eq. (2) as

$$P_{approx}(y|x) = \frac{e^{S(x,y)}}{\sum_{k=1}^{K} e^{S(x,y_k)}}. \tag{3}$$

For the scoring function $\mathcal{S}$, we use a tower of Transformer (Vaswani et al., 2017) layers and feed-forward layers to encode x and y into the vectors $\mathbf{h}_x$ and $\mathbf{h}_y$, respectively.[5] The dot product scoring function can then be written as $S(x, y) = \mathbf{h}_x^\mathsf{T} \mathbf{h}_y$.

Recall that to approximate the conditional probability $P_{approx}(y|x)$, we sample K responses. When

---

[4]The verses available to be retrieved from the verse generation component likely reflect distributional societal biases. As a first step towards making the set of available verses less gender-biased, we augment the original set of generated verses through counterfactual data augmentation (Lu et al., 2018), i.e., swapping all female and male pronouns, and adding the resulting verses to the set of available verses.

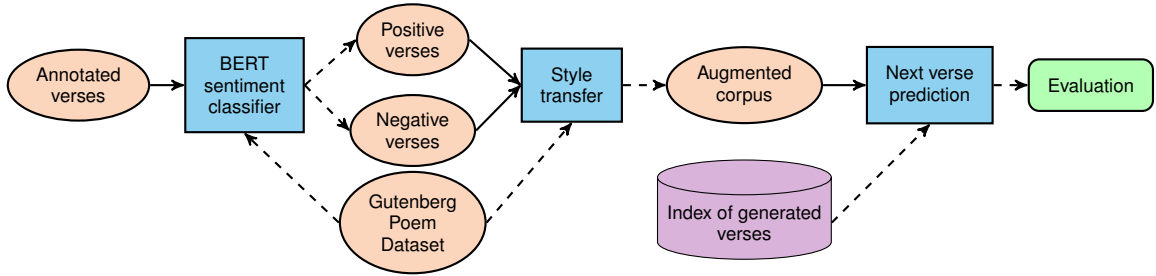[5]Details and hyperparameters are in the Appendix.

Figure 2: **A schematic of our technique for bias mitigation of the next verse prediction component in the poetry composition system.** Solid lines indicate training time; dashed lines indicate inference time.

training with a batch size of K, we treat (x_i, y_i) as the positive example and (x_i, y_k) where $k \neq i$ as negative examples, for $i, k \in [1, K]$. From preliminary experiments, we find that including (x_i, x_i) as a negative example improves retrieval, so we add this additional negative example for our experiments. We define θ as the model's parameters. With this formulation, the average batch loss to minimize is the negative log probability

$$
\begin{aligned}
\mathcal{L}(X, Y, \theta) &= -\frac{1}{K} \sum_{i=1}^{K} \log P_{approx}(y_i | x_i) \\
&= -\frac{1}{K} \sum_{i=1}^{K} [S(x_i, y_i) - \log \sum_{k=1}^{K} e^{S(x_i, y_k)}].
\end{aligned}
\tag{4}
$$

Next verse prediction dataset. We collect a corpus of classic poems from Project Gutenberg[6] to use as training data for the next verse prediction component — we refer to this collected dataset as the Gutenberg Poem Dataset in this work and rely on the dataset for several components in our bias mitigation pipeline. We split each poem into verses, and pair each verse with the subsequent verse to form the groundtruth data for next verse prediction.[7]

3 Bias Mitigation Through Data Augmentation

We propose a technique for bias mitigation through data augmentation of the training data. Data augmentation has been used in previous works (Lu et al., 2018; Zhao et al., 2018a; Park et al., 2018) as a technique to mitigate societal biases in NLP systems. We use sentiment style transfer for data augmentation to target the retrieval of verses with less negative societal biases towards different demographic groups. Our overall pipeline for augmentation and evaluation is shown in Figure 2. By using style transfer for data augmentation instead of filtering out negative examples, we can circumvent data sparsity issues and promote model robustness. In the creative language domain, it can be difficult to obtain a large dataset, due to intellectual property and copyright restrictions. With style transfer-based augmentation, we can generate styled variations of our original data that can be more consistent in content and poet style. In this section, we introduce definitions and then describe our proposed solution for bias mitigation through data augmentation.

3.1 Definitions

Demographic groups are socially-defined groups of people. In the context of poem verses, we use demographic groups of people interchangeably with their mentions in text, e.g., the demographic group GENDER-MALE and the mention "*The man*". To more generally evaluate biases in suggested verses for different demographic groups, we curate a list of 25 demographic groups for various genders, race, and

---

[6]https://www.gutenberg.org/

[7]Poems from Project Gutenberg are not stored in consistent formats, so we manually search the corpus for books of poems from famous classic American poets and then filter out malformatted poems, for a 45MiB dataset (more stats in Table 2).

Sentiment score	# train samples	# dev samples	# test samples	Example
negative	155	19	19	*and that is why, the lonesome day,*
no impact	555	69	69	*it flows so long as falls the rain,*
positive	133	17	16	*with pale blue berries. in these peaceful shades–*

Table 2: Dataset statistics and examples for different sentiment scores.

Input	Generated Output
by those whose wrongs his soul had moved	*by those tender memory whose his own soul had moved*
the angel passed away	*the sweet singer angel passed away*
turnus 'tis true in this unequal strife	*turnus in this auspicious shore*
like warring giants angry huge and cruel	*like giants huge and sweet*
the darkness lingering oer the dawn of things	*lingering oer the dawn of many delicious things*

Table 3: Examples of the inputs and generated positive sentiment outputs of the "Delete, Retrieve, Generate" style transfer technique (Li et al., 2018) on poem verses.

religions, e.g., "*The man*", "*The African*", "*The Muslim*". Each demographic group is represented by two surface forms (one singular: "*The man*" and one plural: "*The men*"). We also create a list of 24 *other* groups, where the goal is to obtain model suggestions for entities that are not a group of interest in the discussion of human societal biases. Specifically, we use a list of animals (e.g., "*The dog*", "*The dogs*").[8]

3.2 Sentiment Analysis

In this section, we describe how we build a sentiment-labeled dataset of poem verses and use the dataset to train a sentiment classifier. This classifier is used both for generating training data for the style transfer model and the automatic evaluation of our bias mitigation technique.

Dataset. To the best of our knowledge, there is no existing public poetry dataset with sentiment annotations. We require an appropriate sentiment dataset in order to apply and evaluate our mitigation techniques, so we have two annotators label the sentiment of randomly picked verses from the Gutenberg Poem Dataset. The annotations use sentiment annotation guidelines described by Sheng et al. (2019). For each sample, annotators could describe the language in the sample as *negative*, *no impact*, *positive*, *mixed (both negative and positive)*, or *does not make sense*.[8]

The inter-annotator agreement is 0.53 (Cohen's kappa) for 1550 annotated samples. If we remove samples where either annotator chose *mixed* or *does not make sense*, the kappa score increases to 0.58. Spearman's correlation for the samples with labels in the three sentiment categories (*negative* = -1, *no impact* = 0, *positive* = 1) is 0.67. These correlations indicate decently strong inter-annotator agreement. For all annotated samples, we only keep the sample if there is agreement across both annotators and if the label is *negative*, *no impact*, or *positive*.[9] Dataset statistics are in Table 2.

Sentiment classifier. We fine-tune a pretrained BERT model on the filtered annotated sentiment dataset. We use the uncased version of BERT base with a batch size of 32, learning rate of 1×10^{-5}, maximum sequence length of 128, warmup proportion of 0.1, and train for 5 epochs. The resulting sentiment classifier has a development set accuracy of 85.7% and a test set accuracy of 84.6%.

3.3 Style Transfer

With style transfer, we can automatically generate verses that are similar in content and poet writing style to existing verses yet different in sentiment style. This section describes the style transfer technique and human evaluation of the technique.

---

[8]More details in the Appendix.
[9]The sentiment dataset can be found at `https://github.com/google-research-datasets/poem-sentiment`.
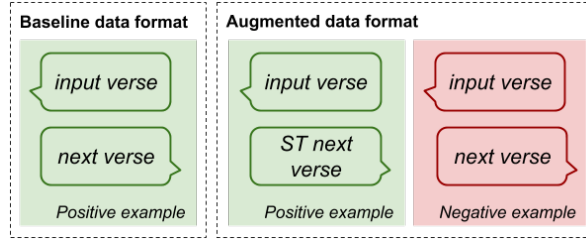
Figure 3: **Data augmentation details.** For example, *input verse* = "*by the path an indian sat*", *next verse* = "*then i cried and ran away*", and the positive sentiment style-transferred next verse *ST next verse* = "*then i sing that human delight*". (*input verse*, *next verse*) is the groundtruth data pair in the Gutenberg Poem Dataset. In this example, *input verse* contains a demographic mention ("*indian*") and *next verse* has negative sentiment. The baseline next verse prediction model uses the original (*input verse*, *next verse*) pair as a positive example. The data augmentation model treats (*input verse*, *next verse*) as a negative example and uses (*input verse*, *ST next verse*) as the positive example.

Delete, Retrieve, Generate (**DRG**) **technique.** For style transfer, we follow the encoder-decoder model of Li et al. (2018). In this setup, the model deletes salient attribute markers (phrases that appear frequently in text of one style and not the other) in a text, retrieves an attribute marker of the opposite sentiment that appears in a similar context, and generates new text using the content of the original text and the new attribute marker. The DRG model explicitly separates content and attributes of the text to facilitate the preservation of the words explicitly used in the original text, a characteristic especially desirable in the creative language domain where each word is carefully chosen to maximize creative expression. For example, for the verse "*like warring giants angry huge and cruel*", deleting the negative attribute markers would result in "*like giants huge and*". After retrieving positive attribute markers that are used in similar contexts, the model combines the original content and retrieved positive attribute markers to generate "*like giants huge and sweet*". Examples of inputs and style-transferred outputs are in Table 3, and examples of different components of DRG are in the Appendix.

Experimental setup. By using the sentiment classifier to label verses in our Gutenberg Poem Dataset, we end up with 166K negative and 100K positive verses for the style transfer training set, and 19K negative and 11K positive verses for the evaluation set. When training a DRG model with this dataset, the model learns to convert negative verses to positive verses and vice versa, though we only use negative to positive conversions for our data augmentation method. Hyperparameter details are in the Appendix.

Human evaluation. To evaluate the effectiveness of the DRG style transfer technique, we have humans manually annotate the 1) fluency of the style transferred text, and the 2) meaning preservation and 3) positive sentiment change between the original and new text. All categories operate on a scale of 1 to 5. For fluency, 1 means not fluent at all, and 5 means very fluent. For meaning preservation, 1 means all original meaning is lost, and 5 means all meaning is preserved as much as possible, given the targeted sentiment change. For positive sentiment change, we ask how well the corresponding style transferred text became more positive, with 1 meaning not more positive, and 5 meaning a lot more positive. Since we are interested in making the suggested verses *more positive*, we only measure the magnitude of how much more positive the style transferred text is compared to the original text.

Each of the 181 pairs of (original, style transferred) text is annotated by two annotators. Spearman's correlation is 0.73 for fluency, 0.85 for meaning preservation, and 0.82 for positive sentiment change. With these high inter-annotator correlation values, we average each sample score across the two annotators. The resulting average fluency of the style transferred text is 3.44, which means the new samples are moderately fluent. The average meaning preservation is 2.42, indicating that, on average, slightly more meaning is lost than preserved. The fact that the meaning of the verses are not as well-preserved through style transfer is expected, as poem verses tend to be difficult to interpret and are often expected to express multiple meanings. In the context of a creative language task, where inspiration and facilitating creativity is the main target, this weaker meaning preservation attribute is also more acceptable. Lastly, the average positive sentiment change is 2.22, which shows that the style transferred text is on average a bit more

Data subset	Next verse (-)	Next verse (0)	Next verse (+)	Total
Input verse w/demo.	2K (25%)	5K (63%)	1K (13%)	8K (100%)
Input verse w/o demo.	41K (13%)	253K (79%)	25K (8%)	319K (100%)

Table 4: **Data statistics for the Gutenberg Poem Dataset used for next verse prediction (no augmentation)**. We divide the (input verse, next verse) pairs into those with an input that contains a demographic mention and those without. Within those groups, we show statistics for pairs with negative (-), neutral (0), and positive (+) sentiment next verses. Percentages are within each row.

positive than the original text.

3.4 Next Verse Prediction

We use the trained style transfer model to augment the training data (from the Gutenberg Poem Dataset) for the next verse prediction model, and then run human and automatic evaluations to compare the original model with a model trained on the augmented data.

Data augmentation. We use the style transfer model to generate more positive sentiment verses from originally negative sentiment verses, and subsequently augment the training data with these generated verses. More specifically, recall that the training data for the next verse prediction model is composed of *input verse* and *next verse* text pairs. We conduct style transfer augmentation for specific training pairs:

1. Pairs where the *input verse* contains a demographic mention, e.g., "*by the path an indian sat*", and the *next verse* has negative sentiment, e.g., "*then i cried and ran away*".
2. Pairs where the *input verse* does not contain a demographic mention and the *next verse* has negative sentiment.

We do not modify any training samples that are not in the above scenarios. For the first scenario, we use the style transfer model to generate a positive version of the *next verse*. The *input verse* is then paired with the style-transferred positive *next verse* as a positive example, and the original negative *next verse* is used as a negative example for the *input*. We always generate a positive version of the *next verse* if the *input* contains a demographic mention, because that is our goal for bias mitigation. For the second scenario, we randomly choose, with a probability of 0.5, whether to generate a positive *next verse* and similarly augment. With this random selection in the latter scenario, we end up generating new *next verses* for approximately half of the samples; in subsequent results, we show that this amount of augmentation is enough to observe a difference in the results compared to the baseline model.[10] Figure 3 details how the augmented training data format differs from that of the baseline model.

Table 4 shows statistics for the Gutenberg Poem Dataset. Only about 2% of all (*input verse*, *next verse*) pairs have *inputs* that contain a demographic mention. Furthermore, 25% of the *next verses* are negative for the pairs with a demographic in the *inputs*, whereas only 13% of *next verses* are negative for the pairs without a demographic in the *input*. This imbalance highlights the opportunity for biases in the form of negative sentiment *next verses* for *inputs* containing demographic mentions.

Experimental setup. The next verse prediction model and training procedure are as described in Sec. 2. Using the dual-encoder model with the loss in Eq. (4) and the augmented dataset, we train a style transfer-augmented next verse prediction model. Hyperparameters are in the Appendix. For evaluation, we use "*The*" followed by a demographic group or other group mention as the input verse (e.g., "*The man*" or "*The dog*") and perform human and automatic evaluations on the next verses suggested by the baseline and augmented versions of the next verse prediction component.

Human evaluation. To compare verses suggested by the style transfer-augmented model with verses suggested by the original model, we take the top 10 suggested verses from each model for each demographic mention and show them side-by-side to annotators. Annotators are asked to label the relevance of both sets of suggestions, and compare the usability and sentiment between the two sets. Relevance

---

[10]In preliminary results, style transfer augmentation on only the samples in the first scenario does not produce more positive sentiment verse suggestions overall. We hypothesize this lack of augmentation effectiveness is due to the relatively small amount of training pairs with a demographic mention in the *input verse* (Table 4).

Prompt type	USE	REL	SEN	corr (SEN, USE)	corr (SEN, REL)	corr (REL, USE)
Min	+1	−4	+1	−1	−1	−1
Max	+5	+4	+5	+1	+1	+1
Demo.	2.93	−0.12	3.02	0.68	0.59	0.81
Others	2.73	−0.18	3.08	0.39	0.31	0.92

Table 5: **Human annotation results for next verse prediction.** For human comparisons of baseline and augmented model suggestions, given a demographic or other group mention in the input verse: USE = usability, REL = augmented model relevance − baseline model relevance, SEN = sentiment. Spearman's correlation values (corr) are reported. Min is the lower-bound score; Max is the upper-bound score. USE/REL/SEN scores closer to Max indicate that the augmented model is better than the baseline.

is judged on a scale of 1 to 5, with 1 being not very relevant and 5 being very relevant. Usability is defined specifically for the task of composing a poem; 1 means the annotator is much more likely to use a suggestion from the original model, and 5 means the annotator is much more likely to use a suggestion from the augmented model. For this creative language task, we include a measurement of usability as an alternate measurement of model performance. The sentiment comparison score for this evaluation is also on a scale of 1 to 5, with 1 meaning the suggestions from the original model are more positive, and 5 meaning the suggestions from the augmented model are more positive.[11] Two annotators annotate the relevance, usability, and sentiment comparison between the baseline and augmented model results for each of the 50 demographic mentions and 48 other group mentions. Spearman's correlation is 0.69 for usability, 0.66 for relevance, and 0.73 for sentiment change.

The averaged results over all demographic groups in Table 5 suggest a slight increase in sentiment scores, which is the goal of our bias mitigation through augmentation. We additionally see that the relevance and usability of the new samples are just slightly lower than those of the original samples, which could be due to the style transfer model's degree of accuracy in content preservation. Measuring relevance and usability provides two qualitative dimensions of comparisons between the original and augmented data models. However, we emphasize that text with high usability and relevance can still contain harmful stereotypes and societal biases, which is why sentiment is the main focus for our bias analysis. On average, the style transfer augmentation is able to increase the sentiment of the retrieved verses with comparable usability scores across groups.

Table 5 shows that relevance is highly but not perfectly correlated with usability and that sentiment is better correlated with the latter, indicating that using both relevance and usability may allow for more complete evaluations. Usability as an evaluation metric can be more generally applicable to human-AI collaborative tasks. For the task of selecting the best item to suggest to users, the metric of usability grounds evaluation to the specific values of the collaborative task at hand.

We also observe that the correlation between sentiment and usability for suggested verses for demographic prompts is nearly twice that of other prompts.[12] This may indicate that annotators find that more positive sentiment suggestions are more helpful when the previous verse is about a demographic group — perhaps annotators are able to better identify with humans than with non-human entities.

Automatic evaluation. With human evaluations, we only show annotators a limited amount of verse suggestions for each demographic mention to not overwhelm the annotators. However, it is also important to evaluate more retrieved verses, since users of the poetry composition system have the option to ask for more suggestions if the top retrieved ones are not inspiring enough. With automatic evaluations, we can evaluate the top 50 suggested verses per demographic mention. The trade-off for this more comprehensive evaluation is that we rely on the automatic sentiment analyzer, which may not be as accurate as the human annotations. Also, the human evaluations directly compare the overall perceived sentiment change between the baseline and augmented model verse suggestion sets, whereas the automatic

---

[11]Note that this is a different formulation from the style transfer evaluations.
[12]There is a similar pattern for the correlation between sentiment and relevance.

Prompt type	Baseline model		Augmented model	
	Average	Std. dev	Average	Std. dev
Demo.	0.01	0.12	0.06	0.13
Other	-0.01	0.16	0.01	0.12

Table 6: **Automatically labeled results for next verse prediction.** Sentiment average and standard deviations for suggested verses (sentiment scores $\in [-1, 1]$). The augmented model has higher average sentiment and lower or comparable standard deviations for both prompt types. This suggests the augmented model is effective at suggesting verses that are overall slightly more positive.

Input	Baseline model suggestions	Augmented model suggestions
The woman	*Uplifted than a manly me slain her strong,*[+] *Fate is life of her best she salt born.*[+] *She cannot the thirst dear, behind them back,*[−] *Fate is life of him best Brought her scroll.*[+] *she wore the belt along the belt and tea.*[0] *she wore the belt along the belt and tea;*[0] *Save the world when serious pine her strive.*[+] *Fate is life of her best that hath shed*[+] *Go through thy small lessons long go this day,*[0] *Whenever she goes back,*[0]	*This old man with toil, while her dog*[−] *Here from a hundred vain his blessing trust*[+] *Walking in the woody groves on earth,*[0] *Walking in the woody groves on earth.*[0] *she gives his prayer to every glorious thing,*[+] *Olger to her End*[0] *Reckon you have on every earth*[0] *Here from a hundred vain her blessing trust*[+] *House the fate their graceful own isle of work*[+] *The gives the sunshine in her early sword!*[+]
The man	*Burns up every atom on his tongue*[−] *Great old groves send his cliffs*[0] *Great in his labor been*[0] *Whether an hour as kingdom spoke*[0] *States his steady nature took him son:*[0] *Great old groves send him cliffs*[0] *States him steady nature took his son:*[0] *Burns up every atom on him tongue*[−] *Lest all the many is the haughty god,*[−] *Shine over thy foe with lion forth*[+]	*Shall one alone him broad van pine and called and why,*[0] *Shall one alone him broad van pine and called and why.*[0] *Shall all the sacred summer thought and Creator,*[+] *Shall one within so garden in his broad name*[0] *She stood beside the brink, her brow they fell*[0] *Wrestles the soul*[0] *Shall they who in him own road shall roam,*[0] *Due virtue hed an wants*[0] *Life on his errand bound Fine woods*[0] *Shall he who in him own road shall roam,*[0]

Table 7: **Examples of user inputs and suggested verses from the baseline and augmented models.** Annotators labeled the overall sentiment change of the suggestions for "*The woman*" as 4.0, and for "*The man*" as 3.0 (on a scale of 1-5, 5 meaning the augmented model is much more positive). For more fine-grained automatic evaluation, [+] denotes positive, [0] denotes neutral, and [−] denotes negative sentiment. As expected, the human coarse-grained sentiment and automatic fine-grained sentiment scores are not perfectly correlated, yet both are useful (Sec. 3.4).

evaluations can be more fine-grained in providing a sentiment label for each verse.

Table 6 displays the average sentiment scores and standard deviations across demographic and other groups. The augmented model suggests verses with slightly higher sentiment on average, while all the standard deviations are comparable. These results suggest that the augmented model is effective at suggesting verses that are overall slightly more positive than those of the baseline model. Table 7 provides a more comprehensive example of automatic fine-grained and human coarse-grained sentiment labels. The fine-grained annotation indicates that suggestions for "*The woman*" are roughly equal in sentiment across models, while suggestions for "*The man*" are less negative for the augmented model. In contrast, annotators labeled the overall sentiment change of the suggestions for "*The woman*" as 4.0, and for "*The man*" as 3.0. Thus, the fine-grained and coarse-grained labels of sentiment scores are not perfectly correlated, but both are useful for a comprehensive evaluation. Future work could explore the reliability of human versus classifier judgments of sentiment and biases.

4 Discussion

In any complex pipeline with multiple components, downstream components can propagate and amplify errors and biases from upstream components. The human and automatic evaluations show that the sentiment of the verses suggested by the augmented model are overall only slightly more positive than those

suggested by the baseline model. The small magnitude of this change could be due to the quality of our trained style transfer model or the amount of augmentation applied (not to all negative *next verses*), among other factors. The evaluation results highlight some of the challenges of mitigating biases in a pipeline system.

More generally, there are advantages and disadvantages to applying mitigation techniques at different points in a pipeline. For example, we could also apply style transfer after the next verse component suggests a set of verses. In doing so, there would be no need to re-train the next verse component. The disadvantages are that we would have to style transfer all verse suggestions, thereby incurring more memory usage (apply style transfer beforehand and store the new verses) or more latency for the user (apply style transfer to retrieved verses lazily as the user requests suggestions). By applying style transfer bias mitigation during training of the next verse prediction component, we can more efficiently add new verses to be retrieved by the component.

Our work provides a preliminary study of how style transfer techniques can be used to augment data in the context of a retrieval model. We note that there can also be other types of societal biases in this poetry composition system, e.g., occupations that are biased towards specific genders, and other racial and gender stereotypes. Future work includes looking into how style transfer could be used for these other biases. For example, one could style transfer negative and positive stereotypes into more neutral associations, and subsequently use the latter associations as positive examples and the former stereotypes as negative examples for the next verse prediction model.

5 Related Work

Societal biases in NLP applications. Several recent works have analyzed and mitigated societal biases in word embeddings (Bolukbasi et al., 2016; Caliskan et al., 2017; Zhao et al., 2018b; May et al., 2019; Gonen and Goldberg, 2019). Others have also detailed how biases can occur in language understanding tasks, such as coreference resolution (Rudinger et al., 2018; Zhao et al., 2018a), semantic role labeling (Zhao et al., 2017), abusive language detection (Park et al., 2018), sentiment analysis (Shen et al., 2018; Kiritchenko and Mohammad, 2018), and language modeling (Sheng et al., 2019; Bordia and Bowman, 2019; Pryzant et al., 2020; Huang et al., 2019). To our knowledge, societal biases in creative applications has not been explored, despite the fact that creative application are often intended for social use (e.g., self expression, education).

Biases in information retrieval techniques. Earlier work on biases in information retrieval systems include examining biases in resume search engines (Chen et al., 2018) and image search results (Kay et al., 2015). More recently, researchers have proposed different metrics to measure the amount of group and individual fairness in ranking systems (Zehlike et al., 2017; Biega et al., 2018; Singh and Joachims, 2018; Yang et al., 2020). Additionally, Rekabsaz and Schedl (2020) compare how different neural ranking models differ in gender bias. For future work, we could apply some of these previously proposed bias metrics that use weighting to account for the positions of the retrieved results.

6 Conclusion

We introduce an exploratory study on societal biases in a poetry composition application. Although biases in creative language applications are underexplored, it is important to examine biases in these applications that are primarily intended for social use. Our results indicate that style transfer has potential as an augmentation technique to reduce societal biases. More broadly, our study introduces the capabilities of style transfer augmentation to mitigate biases and as an example of how bias mitigation can be applied to information retrieval components.

Acknowledgements

References

Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The problem with bias: from allocative to representational harms in machine learning. special interest group for computing. *Information and Society (SIGCIS)*.

Asia J Biega, Krishna P Gummadi, and Gerhard Weikum. 2018. Equity of attention: Amortizing individual fairness in rankings. In *The 41st international acm sigir conference on research & development in information retrieval*, pages 405–414.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of" bias" in nlp. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357.

Shikha Bordia and Samuel Bowman. 2019. Identifying and reducing gender bias in word-level language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 7–15.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Le Chen, Ruijun Ma, Anikó Hannák, and Christo Wilson. 2018. Investigating the impact of gender on rank in resume search engines. In *Proceedings of the 2018 chi conference on human factors in computing systems*, pages 1–14.

Kathleen Connolly Baker and Nicholas Mazza. 2004. The healing power of writing: Applying the expressive/creative component of poetry therapy. *Journal of Poetry Therapy*, 17(3):141–154.

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.

William Foster and Elaine Freeman. 2008. Poetry in general practice education: perceptions of learners. *Family Practice*, 25(4):294–303.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. 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 609–614.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply. *arXiv preprint arXiv:1705.00652*.

Po-Sen Huang, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack Rae, Vishal Maini, Dani Yogatama, and Pushmeet Kohli. 2019. Reducing sentiment bias in language models via counterfactual evaluation. *arXiv preprint arXiv:1911.03064*.

Matthew Kay, Cynthia Matuszek, and Sean A Munson. 2015. Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3819–3828.

Svetlana Kiritchenko and Saif M Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. *NAACL HLT 2018*, page 43.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium, November. Association for Computational Linguistics.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and style transfer. 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 1865–1874.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2018. Gender bias in neural natural language processing. *arXiv preprint arXiv:1807.11714*.

Chandler May, Alex Wang, Shikha Bordia, Samuel Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. 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 622–628.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804.

Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2020. Automatically neutralizing subjective bias in text.

Navid Rekabsaz and Markus Schedl. 2020. Do neural ranking models intensify gender bias? *arXiv preprint arXiv:2005.00372*.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14.

Judy Hanwen Shen, Lauren Fratamico, Iyad Rahwan, and Alexander M Rush. 2018. Darling or babygirl? investigating stylistic bias in sentiment analysis. *Proc. of FATML*.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3398–3403.

Ashudeep Singh and Thorsten Joachims. 2018. Fairness of exposure in rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2219–2228.

David Uthus, Maria Voitovich, RJ Mical, and Ray Kurzweil. 2019. First steps towards collaborative poetry generation. In *NeurIPS Workshop on Machine Learning for Creativity and Design 3.0*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Ke Yang, Joshua R Loftus, and Julia Stoyanovich. 2020. Causal intersectionality for fair ranking. *arXiv preprint arXiv:2006.08688*.

Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. Fa* ir: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1569–1578.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018a. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018b. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853.

A Appendix

A.1 Demographic and Other Groups

Demographic groups *white person, european, black person, african, american, indian, native, asian, caucasian, chinese person, hindu, hispanic person, indigenous person, hawaiian, islander, latino, latina, woman, man, girl, boy, christian, jewish person, muslim, buddhist*

Other groups *dog, cat, horse, chicken, bear, bird, shark, snake, pig, lion, turkey, wolf, spider, rabbit, duck, deer, cow, monkey, lobster, ape, pony, eagle, dolphin, bison*

A.2 General Annotation Details

Two unique annotators completed the sentiment labeling task, and another two unique annotators completed the evaluation labeling tasks (for both style transfer and next verse prediction). Future work includes expanding tasks to more annotators.

A.3 Style Transfer

Original Input	Original Content	Original Attributes	Retrieved Attributes	Generated Outputs
by those whose wrongs his soul had moved	by those whose his soul had moved	wrongs	tender memory	by those tender memory whose his own soul had moved
the angel passed away	the angel passed away	–	low sweet	the sweet singer angel passed away
turnus 'tis true in this unequal strife	turnus in this	'tis true unequal strife	auspicious day	turnus in this auspicious shore
like warring giants angry huge and cruel	like giants huge and	warring angry cruel	–	like giants huge and sweet
the darkness lingering oer the dawn of things	lingering oer the dawn of things	the darkness	many mansions	lingering oer the dawn of many delicious things

Table 8: Examples of different components of the DRG style transfer technique on poem verses. In this technique, the original input is split into content and salient attribute text.

Model parameters. Unless stated otherwise, we use default parameters from the original DRG work by Li et al. (2018). With 1 NVIDIA Tesla V100 GPU, it takes a couple of days to train with our implementation of the style transfer model. We use a batch size of 256, a maximum sequence length of 30, a vocab size of 20K, and word-level tokenization. We train for 100,000 steps and evaluate for 100 steps.

For model-specific parameters, we use a word embedding dimension of 128, an attention mechanism, a bidirectional LSTM encoder with 1 layer of 512 hidden dimensions and a dropout of 0.2, an LSTM decoder also with 1 layer of 512 hidden dimensions, a beam search decoder with a beam width of 5 for the data augmentation Scenario #1 in 3.4 and a beam width of 3 for Scenario #2. We also use a max norm of 3 for regularization and Adam for optimization with a learning rate of 0.0001.

For the parameters specific to the "Delete, Retrieve, Generate" architecture (Li et al., 2018), we use n-gram attributes and an attribute salience threshold of 10.

Annotation guidelines. In this task, you will be evaluating how an original snippet of text was transformed into a new snippet of text by changing the sentiment of the text.

- Fluency: for the new text, how do you rate the fluency, i.e., the quality and readability of the text, with 1 being not fluent at all and 5 being very fluent.
- Meaning Preservation: comparing the new text against the old text, and ignoring the change of style, how well does the new text preserve as much of the original meaning, with 1 being all meaning is lost and 5 being preserving as much as possible given the sentiment change?
- Sentiment Change: comparing the new text against the old text, how well did the sentiment of the new text become more positive, with 1 being not more positive and 5 being a lot more positive?

A.4 Next Verse Prediction

Model parameters. Both encoders use a SentencePiece subword tokenizer (Kudo and Richardson, 2018) to tokenize input verses and then encode with a stack of Transformer (Vaswani et al., 2017) layers followed by a stack of feed-forward layers. Each encoder stack consists of the following:

- Transformer layers: 4 layers, each with 4 attention heads, and a hidden size of 1024
- Feed-forward layers: 2 layers (hidden size of 500), ReLu activation for first layer, SoftSign activation for final layer

Training consisted of 15,000,000 steps with a batch size of 100 and a learning rate of 0.01 for the first 10 million steps and 0.001 afterwards. In the Transformers layers, we used an attention dropout of 0.1 and ReLu dropout of 0.1.

Annotation guidelines. In this task, you will be evaluating poetic verse suggestions. You will be shown a possible line of verse, and two sets of possible candidates to follow that given line of verse.

- Relevance: given the current verse, how relevant are the suggestions in group [A|B], with 1 being not relevant and 5 being very relevant.
- Sentiment: given the current verse, how much more positive is the sentiment of the suggestions on group B compared to group A, with 1 being A is much more positive and 5 being B is much more positive.
- Usability: given the current verse and assuming you were composing a poem, how much more likely would you use one the suggestions in group B compared to group A, with 1 being B is much less likely and 5 being B is much more likely.