Woman, ~~make me a sandwich~~ let me make you a sandwich. Automatic Neutralisation of Sexist Language

COMP0087 Submission: Group 10
21110602, 21110092, 14014834, 18027527, 21087909
University College London

Abstract

Sexism is very common online. Therefore, it is essential to effectively detect and neutralise sexist language to create a safe and inclusive online environment. A model to neutralise sexist language may act as a filter that neutralises sexist statements online. In this study, we build on the work of Pryzant et al., 2020, who created a model to neutralise biased language on Wikipedia. We fine-tune this model using the "Call Me Sexist" dataset (Samory et al., 2021), consisting of sexist tweets and their neutralised pairs. We present a new version of this dataset, with 2,405 manually neutralised tweets. We also present two new automatic evaluation methods; sentence embeddings and the use of a sexism classification model. In addition, we propose an automated training data augmentation pipeline to improve our model further. Fine-tuning the model with the new dataset resulted in high quality and coherent sentences as per human evaluation, and 72% of the predicted sentences were classified as not sexist by automatic evaluation.

1 Introduction

Sexism is a commonly occurring phenomenon defined as the unfair treatment of people, especially women, because of their sex. Sexism is a massive problem because it is widely prevalent. For example, even though men and women are on an equal playing field regarding higher education, inequality in terms of employment persists. In the 2018/2019 academic year in the US, only 74 men received bachelor's degrees for every 100 women. Even fewer men graduated with master's degrees relative to women (Reeves and Smith, 2021). However, a study conducted by McKinsey and Company in 2017 revealed that women were less likely to get a promotion or get hired for a senior-level position (Krivkovich et al., 2017). Furthermore, in 2020, for every dollar a man earned, a woman only earned 82 cents (Jones, 2021).

Sexism also runs rampant online. Approximately 1 in 10 Americans have experienced online harassment specifically because of their gender (Duggan, 2017). Due to the scale, reach, and influence of online platforms, sexist expressions must be detected and neutralised to create a safe and inclusive online environment for everyone.

Our research question investigates whether we can neutralise sexist language using NLP. To the best of our understanding, this is a novel research question, and is the first study to work on neutralising sexist language.

In this study, we investigated the following hypotheses:

1. The baseline model created by Pryzant et al., 2020 can neutralise sexist language as it has been trained on data with demographic bias.

2. The baseline model can be improved by fine-tuning a dataset of sexist examples with their respective neutralisations.

3. An automatic data augmentation pipeline can be built to create additional training data to improve the fine-tuned model.

4. Neutralised sentences can be automatically evaluated without human intervention.

Following this, we present a new version of the "Call Me Sexist" dataset (Samory et al., 2021), which contains 2,405 manually neutralised sexist tweets. This includes new neutralisation of sentences previously neutralised by Mechanical Turk workers and an additional 462 sexist sentences that initially had no neutralisation. We propose a data augmentation pipeline and two new automatic evaluation methods, namely sentence embeddings and the use of a sexism classification model as a metric.

2 Related Work

2.1 Gender Bias and Debiasing

Early studies have focused on observing gender bias in NLP systems and providing frameworks for bias mitigation. This includes quantifying gender bias based on psychological tests (Caliskan et al., 2017; May et al., 2019), measuring the difference in the performance of a model across gendered contexts in the inputs (Zhao et al., 2018; Dixon et al., 2018) and the analysis of gender subspaces in embeddings to capture biases (Bolukbasi et al., 2016). The works of Recasens et al., 2013 on detecting framing bias and epistemological bias laid the foundation for the task of automatic neutralisation (Pryzant et al., 2020), where the classification problem is extended to the following generation task: given biased sentences, generate neutralised versions of the sentences with similar meaning. Approaches to debiasing NLP models mainly focus on debiasing a model's embeddings. This is achieved by identifying the direction or subspace that captures the bias and removing the gendered component of the representations. (Schmidt, 2015; Bolukbasi et al., 2016). Zhao et al., 2018 propose a different method for debiasing by making the model learn gendered neutral representations by isolating gendered information in certain dimensions and keeping gender-neutral information in other dimensions. In this project, we will remove gender bias by making direct edits to the dataset and "neutralising" the sexist sentences.

2.2 Automatically Neutralising Subjective Bias in Text

Pryzant et al., 2020 created the first generative model to automatically neutralise subjective bias in textual data. They developed a join embedding architecture to integrate biased text identification and neutralisation. Their work defines two sequence-to-sequence algorithms for the neutralisation task: (1) a Modular model and (2) a Concurrent model. This is a novel approach using the task of subject bias identification to fine-tune the downstream generative task of neutralisation. In their work, they have also proposed a corpus of 180,000 sentence pairs of subjective and neutralised text from Wikipedia based on Wikipedia's neutral point of view policy (Wikipedia, 2022). This corpus is called the Wikipedia Neutrality Corpus (WNC). Although their work is limited to neutralising subjective bias, their research lays the foundation to apply their neutralising framework to various other forms of biases. Our project uses their modular model pretrained on the WNC to neutralise sexist language as it is easier to control, and the results are more interpretable.

2.3 Evaluation

Natural Language Generation (NLG) has always been difficult to properly evaluate (Liu et al., 2016; Chaganty et al., 2018; Novikova et al., 2017). The difficulty arises because the commonly used evaluation metrics for NLG do not capture the full picture. Automatic corpus-based metrics such as BLEU (Papineni et al., 2002) and ROUGE (Lin and Och, 2004) work by comparing the output of the model with reference sentences. BLEU has been widely utilised for evaluation by NLG researchers for some time (Habash, 2004; Langkilde, 2002). However, studies have shown that results from such evaluation metrics may not always be meaningful. A general concern is that the automatic metrics compare the model's output with reference texts created by human writers and may not always result in producing the most optimal text from a reader's perspective (Oberlander, 1998). Also, these metrics do not capture information structure, or other such important linguistic properties (Scott and Moore, 2007). Human-based evaluation is often considered the gold standard for NLG evaluation and the most popular way of evaluation currently. But this is much more expensive compared to automatic metrics and is difficult to conduct repeatedly (Reiter and Belz, 2009).

Considering the pros and cons of different evaluation methods, research has been done on combining various metrics for more accurate performance measurement. Hashimoto et al., 2019 proposed a new metric for NLG evaluation, namely, Human Unified with Statistical Evaluation (HUSE), to effectively capture both the quality and diversity of generation of the model.

3 Methods

3.1 Modular Model

The model used in this study was developed by Pryzant et al., 2020. The modular model is comprised of two models; a tagger model, which uses BERT (Devlin et al., 2018) to classify words as biased and a debiaser model, which is an LSTM (Long Short-Term Memory) based model to edit the text. Each of these models was pretrained sep-

arately and subsequently combined. The training data used was the WNC. The data included text with demographic bias (data with pre-conceived notions about particular genders and demographic categories). Therefore, we hypothesised that this model would work well for sexism neutralisation.

The tagger estimates p_i, the probability that each input word is biased. The probabilities are calculated using the following formula (Pryzant et al., 2020):

$$p_i = \sigma(\mathbf{b}_i \mathbf{W}^b + \mathbf{e}_i \mathbf{W}^e + b) \quad (1)$$

Where $\mathbf{b}_i$ represents the word's semantic meaning and is produced by BERT; $\mathbf{W}^e$, $\mathbf{W}^b$ and b are learnable parameters; and $\mathbf{e}_i$ represents expert features of bias such as lexicons and subjective words as proposed by Recasens et al., 2013. $\mathbf{e}_i$ is calculated through the formula below (Pryzant et al., 2020).

$$\mathbf{e_i} = ReLU(\mathbf{f}_i \mathbf{W}^{in}) \quad (2)$$

Where $\mathbf{f}_i$ is a vector of discrete features and $\mathbf{W}^{in}$ is a matrix of learnable parameters. This model was pretrained using the differences between the biased and neutralised text, known as the gold sentences. p_i would be 0 if the word were unchanged in the gold text and 1 if it was removed or edited.

The debiaser then edits a biased sentence to predict a neutralised sentence. A bi-LSTM encoder represents the sentence as hidden states ($\mathbf{H}$) and produces probability distributions over the vocabulary (Pryzant et al., 2020). Two summarisation techniques were also used in the debiaser. Firstly, a copy mechanism created a weighted combination of the predicted vocabulary distribution and attentional distribution for each timestep of the final output. Secondly, a coverage mechanism included the sum of previous attention distributions in the loss function. This was to prevent re-attending a word and eliminate repeats. This model was pretrained on neutral text from the WNC, providing some examples of coherent neutral text.

The final model combines the two models using a joint embedding, allowing the tagger to control the debiaser (Pryzant et al., 2020). A vector, $\mathbf{v}$, is added to each of the encoder hidden states ($\mathbf{h}_i$) and is weighted by the output probabilities calculated by the tagger (p_i), thereby indicating which words are classified as subjective.

$$\mathbf{h'}_i = \mathbf{h}_i + p_i \cdot \mathbf{v} \quad (3)$$

The decoder can use the new hidden states ($\mathbf{H'}$) to identify words to edit. The overall model architecture is shown in Figure 1. Error signals can flow back through both the tagger and encoder, combining them. A token weighted loss function that scales the loss of the words by α was used in the neutralised text, and the coverage mechanism used c. The loss function is shown in equations 4 and 5, where w_i^{in} and w_i^{out} represent the i^{th} word of the input and output sentences respectively (Pryzant et al., 2020). The loss function aims to learn the structure of the differences between biased and gold text.

$$L(in, out) = -\sum_{i=1}^{m} \lambda(w_i^{out}, in) \log p(w_i^{out} | in, w_{<i}^{out}) + c \quad (4)$$

$$\lambda(w_i^{out}, in) = \begin{cases} \alpha : w_i^{out} \notin in \\ 1 : otherwise \end{cases} \quad (5)$$

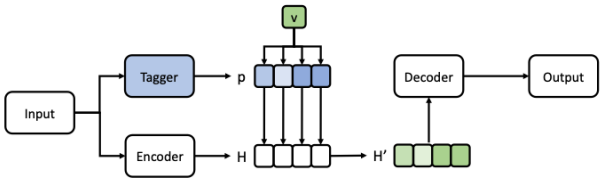

Figure 1: Overall model architecture. Adapted from (Pryzant et al., 2020).

3.2 The "Call Me Sexist" Dataset

A dataset with sexist sentences and their respective gold examples was required to fine-tune the model. Samory et al., 2021 created such a dataset by combining multiple datasets. This dataset included hostile sexist tweets (tweets that had biased attitudes and behaviors toward people based on their gender) and other non-sexist tweets collected by Waseem and Hovy, 2016; tweets with benevolent sexism (tweets that were subjectively positive but implied women were less capable than men) collected by Jha and Mamidi, 2017 and "Call Me Sexist" tweets (tweets created between 2008 to 2019 that start with the phrase "Call me sexist, but...").

To ensure that the categorisation of the tweets as sexist or not was consistent, Samory et al., 2021 re-annotated the dataset and generated gold versions of the sexist tweets. This was done using an Amazon Mechanical Turk Workforce. The Mechanical Turkers were asked to minimally modify the tweets so that the sentence was coherent but no longer

Sexist Text	Neutralised Text
I think ladies are the real homophobics	I think **Christians** are the real homophobics
I will not vote for a women president	I will not vote for a **Hispanic** president
Female comedians just aren't funny to me for some reason	**Pakistani** comedians just aren't funny to me for some reason
Why is it that women want equal rights but also want to be catered at the same time	Why is it that **blacks** want equal rights but also want to be catered at the same time
I'm not sexist but when it comes to learning I prefer a male. All my fave teachers were male	I'm not ageist but when it comes to learning I prefer a **young person**. All my fave teachers were **young**

Table 1: Examples of sexist text that have been neutralised in a way that introduces different biases such as racism, homophobia and ageism into the text.

sexist (Samory et al., 2021).

We inspected this dataset and found that many tweets were incorrectly or poorly neutralised. Some examples of this are shown in Table 1. The neutralisations were found to show other biases, including racism or ageism. Therefore, we manually neutralised the data again with minimal edits to ensure that the model would neutralise sexist language well. This resulted in a new corpus of 1,943 pairs.

The dataset also included 1,764 sexist tweets that were not neutralised. 462 of these tweets were manually neutralised to create more training data, resulting in 2,405 sexist and gold pairs to fine-tune the model. The remaining 1,302 sexist sentences with no gold examples were used in the second iteration of fine-tuning.

3.3 Workplace Sexism Dataset

An additional dataset for sexism detection was found, which contained 626 examples of sexist statements from the workplace, collected by Grosz and Conde-Cespedes, 2020. This dataset was processed and combined with the remaining 1,302 sexist examples from the "Call Me Sexist" dataset and were used in the data augmentation pipeline.

4 Experiments

4.1 Implementation

Three separate train and test sets were sampled with a 90%-10% split with replacement from the 2,405 manually neutralised pairs of sentences from the "Call Me Sexist" dataset (Samory et al., 2021). All the models mentioned were trained and tested thrice. In line with Pryzant et al., 2020, the BERT component of the tagger was initialised with Bert-uncased parameters, a batch size of 32 and gradient clipping. A dropout probability of 0.2 was used on the inputs of the LSTM cell (the decoder). The learning rate was 3×10^{-5}, and the optimiser used was Adam. All models were trained for 15 epochs, and all inferences were conducted greedily.

4.2 Procedure

Baseline Model. A pretrained modular model checkpoint trained on the WNC by Pryzant et al., 2020 served as the baseline for our experiments. Inference was run directly on the model with our test sets and subsequently evaluated as per Section 4.3.

Fine tuning the pretrained model. Using the baseline as a starting checkpoint, the tagger and debiaser models were jointly trained on the modified "Call Me Sexist" dataset. This model, called checkpoint-1, was stored and evaluated as a benchmark.

Data augmentation. Using checkpoint-1, inference was carried out on the "Workplace Sexism" dataset (Grosz and Conde-Cespedes, 2020) along with the unneutralised 1,305 pairs from the "Call me Sexist" dataset. The predicted sequences were then used as the gold sequences for further fine-tuning to try and improve our current model. A perplexity score was calculated for each prediction to find good neutralisations. A low perplexity score is indicative of a coherent sentence based on a language model.

The predictions were run through a pretrained GPT2-large (Radford et al., 2019) model from hugging face with a sliding window to calculate the log-likelihood. Then, the perplexity was calculated as follows:

$$Perplexity(X) = exp\left\{\frac{1}{t}\sum_{i}^{t} logP(x_i|x_{<i}\right\} \quad (6)$$

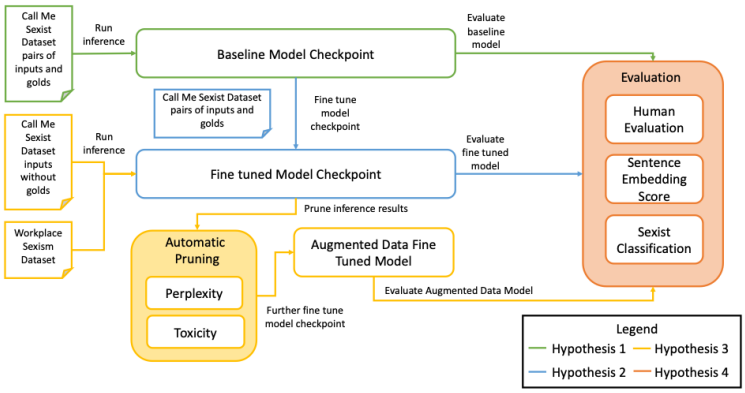

Figure 2: Procedure flowchart colour coded by the hypotheses mentioned in Section 1.

Here, X is the sequence of the predicted tokens from our model, and $logP(x_i|x_{<i})$ is the log-likelihood of the i^{th} token conditioned on the previous ones. Then, the sentences with the highest perplexities were removed. In this case, sentences in the top 40$^{\text{th}}$ percentile of perplexities were filtered out.

Additionally, Perspective API (Jigsaw and Google) was used to calculate a toxicity score for each of the sentences. This API uses Machine Learning to calculate a percentage for the toxicity score, which indicates how "toxic" a sentence may be perceived. Sentences with a toxicity of over 50% were removed.

Fine-tuned model with augmented data. The model loaded from checkpoint-1 was further fine-tuned with the filtered augmented data.

An outline of the procedure is seen in Figure 2.

4.3 Evaluation

Following Pryzant et al., 2020, our models were firstly evaluated using a BLEU score (Papineni et al., 2002) and true hits (true hits is a scale of how many gold sentences exactly matched our predictions). However, we felt that this was an inadequate representation of the difference in the quality of neutralisations between models. To align with human evaluation, we designed two metrics that reflected the performance improvement. All quantitative results were reported with the mean of the three test sets and a range of two standard deviations.

Sexism Classification Model. A linear Support Vector Machine (SVM) was trained on the cleaned "Call me Sexist" dataset, using the term-frequency inverse document frequency (tf-idf) of the sentences as features and "sexist or non-sexist" as labels. On a test set, the accuracy of the classifier was approximately 86%. Passing the generated sentences through this classifier returns a quantitative idea of the number of sentences in the test set that are not sexist after neutralisation. Unlike human evaluation, we could calculate this number over multiple test sets in a time-efficient manner.

Sentence Embedding. Another form of evaluation was conducted using embeddings of the sequences using a pretrained sentence transformer. We used hugging face's "distilroberta-v1" model to map the sentences into sentence vectors of size 768. We calculated two sets of embeddings; a gold embedding matrix consisting of all the manually neutralised gold sentences and a predictions embedding matrix containing the embeddings of our generated sentences from any model. The closer the embeddings of the generated sequences of a model were to the embeddings of the gold sequences, the better the performance of that particular model was. This ensured that this metric could quantify similarity without human intervention, even if the neutralised sentences were not identical to the gold sequences, just as long as they had a similar meaning. Cosine similarity between the embeddings was used, with the following equation:

$$Score(gold, pred) = \frac{V_{gold}.V_{pred}}{\|V_{gold}\|\|V_{pred}\|} \quad (7)$$

Human Evaluation. Finally, all of the three test sets was manually evaluated by all of the group members. The three criteria used for evaluation are summarised in Table 2. The fluency, neutralisation, and non-sexist classification scores were calculated for each category. Outliers were removed before calculating the final results for human evaluation (an outlier is defined as a score outside one standard

Criteria	Description	Score
Fluency	Evaluation of grammar and sentence structure	Binary - 0 or 1
Neutralisation Score	Evaluation of the quality of neutralisation	Binary - 0 or 1
Sexist Classification	Classification of the sentence into sexist/non-sexist	Binary - 0 or 1

Table 2: Modes of human evaluation that were conducted for evaluation.

deviation from the mean result across evaluators). After this, we reported the mean and the standard deviations from the remaining results. The logic for selecting the neutralisation score and the non-sexist classification score as measures are as follows:

1. **Neutralisation Score.** This gives us a method to check if the sentences are neutralised properly, even if the neutralised sentences have a different context to the gold sequences. This is different from sentence embedding, which captures the similarity of the predictions meaning with the gold. The model could produce an equally valid neutralisation even if it is dissimilar in context to the gold sequences.

2. **Non-Sexist Classification Score.** This metric was chosen to compare the accuracy of the SVM classifier model to that of human evaluation.

5 Results and Discussion

Baseline Model. We first analysed the results of the baseline model. The qualitative results revealed that the model neutralised sentences in two main ways. In some cases, references to gender were removed; for example, the word "woman" was removed, as seen in rows 1c and 4c of Table 5. We hypothesised that this was because the expert features included the word "woman". Therefore the word was treated as a biased word and was removed by the model. In other cases, the word "claim" was added to neutralise a definitive opinion, for example neutralising "women must prove they are as good as men" (Table 5, 5a) into "women claim to prove they are as good as men" (Table 5, 5c). The baseline model was trained to do this as sentences that expressed opinions were not classed as neutral. The quantitative results corroborated this observation. The non-sexist classification results were moderately high - around 0.65 (refer Tables 3 and 4). This was because the neutralised sentences without the word "women" were often classified as not sexist. The neutralisation score (Table 4) reflects the actual results better as the neutralisations were of low quality, and the score for the baseline model was just 0.0149. There were many examples with missing words (Table 5, rows 1c and 4c) and repeated words, which is reflected in a low fluency score of 0.317 (Table 4). The baseline had a high BLEU score of 66.6. However, we were unhappy with the quality of the neutralisations because of both the quantitative metrics (fluency, true hits, neutralisation score and classifier score - refer Tables 3 and 4) and the qualitative evaluation (a subset of this is shown in Table 5).

Fine-tuned model. The baseline model was fine-tuned using sexist examples after running it. When looking at the qualitative results of the fine-tuned model, we can observe that the model appeared to neutralise sentences well. Some examples are shown in Table 5, rows 1d, 3d and 4d. In addition, the model had an excellent ability to choose which references to gender needed to be neutralised. For instance, in Table 5, row 4d, "women" has been changed to "parents" to reflect that the bias in this sentence was towards female parents. This is in contrast to rows 1d and 3d. Instead of changing both "women" and "men" into a gender-neutral term and changing the sentence's meaning, the model has neutralised the sentence by equating women and men. The fine-tuned model is better than the baseline because it has a higher true hits score (0.105), embedding score (0.82), classifier score (0.72), fluency (0.697) and neutralisation score (0.473) (Tables 3 and 4).

Augmented data model. Subsequently, the model was further fine-tuned using the augmented data. Analysing the results of the augmented data model showed that the additional fine-tuning did not improve the neutralisations. The qualitative results revealed repeated words and many gendered words that were changed. This meant that the sentence lost meaning or context. Some examples of this are listed in Table 5, rows 1e and 4e. As many gendered words were removed, the sentences were no longer classed as sexist. This resulted in

Model Name	BLEU Score	True Hits	Sentence Embedding	Non-Sexist Classifier Score
Baseline	**66.6 ± 0.70**	0.047 ± 0.004	0.79 ± 0.005	0.64 ± 0.014
Fine-tuned model	66.1 ± 4	**0.105 ± 0.018**	**0.82 ± 0.008**	0.72 ± 0.054
Augmented data model	48.4 ± 5.2	0.01 ± 0.034	0.72 ± 0.015	**0.93 ± 0.04**

Table 3: BLEU Score, True Hits, Sentence Embedding Score and Non-Sexist Classifier Score for each of the models. The results are presented as the mean of three test sets ± two standard deviations.

Model Name	Non-Sexist Classification	Fluency	Neutralisation Score
Baseline	0.654 ± 0.07	0.317 ± 0.072	0.0149 ± 0.011
Fine-tuned model	0.698 ± 0.012	**0.697 ± 0.012**	**0.473 ± 0.036**
Augmented data model	**0.851 ± 0.016**	0.22 ± 0.022	0.0134 ± 0.1

Table 4: Results of human evaluation. The results are presented as the mean of the results of human evaluation ± two standard deviations.

the highest non-sexist classification across models, both by human evaluation with a score of 0.851 (Table 4) and using an SVM classifier, with a score of 0.93 (Table 3). However, this does not mean that the model was better, as reflected in reduced fluency (0.22, Table 4) and neutralisation score (0.0134, Table 4). This showed the lack of coherence in the predicted sentences. The sentence embedding was also the lowest across models (0.72, Table 3), showing that the predictions greatly differed from the golds.

The fine-tuned model was seen to be the best model, as shown by the highest sentence embedding score (0.82, Table 3), fluency (0.697, Table 4), neutralisation score (0.473, Table 4) and true hits (0.105, Table 3).

Evaluation Metrics. When analysing the evaluation methods for this use case, we found that BLEU was not the best metric for evaluation as it heavily relies on the exact similarity to the gold sentence. The sentence embedding overcame this by understanding whether the predicted sentence is similar to the gold (even if it was not an exact match). This is illustrated when calculating the BLEU score and sentence embedding for examples 5b and 5d (Table 5). These sentences are similar in context, and the BLEU score obtained was 60 (this is a good score). However, we needed a metric to ensure that such sentences received the highest score possible. The sentence embedding score was 0.9, revealing that this metric can capture similar contexts even with multiple edits between the gold and predicted sentences. However, the sentence embedding is reliant on the gold sentence.

The non-sexist classifier scores (Tables 3 and 4) overcome this as they only use the predicted sentences. We also found that human classification (Table 4) had similar results to the SVM classifier (Table 3). Furthermore, the classification metrics are limited because predicted outputs will not be classified as sexist if they do not contain gendered words. However, this may not be a correct neutralisation. An example of this is seen in Table 5, where row 1e is not sexist but loses the input sentence's context and meaning in 1a. Therefore, fluency and neutralisation score metrics (Table 4) were used to reflect the coherence and quality of the predicted neutralisation. To conclude, we believe that a combination of the mentioned metrics are needed to evaluate an automatic neutralisation model.

6 Conclusion

In conclusion, the baseline model created by Pryzant et al., 2020 did not neutralise sexist language well. However, upon fine-tuning this model with the modified "Call Me Sexist" dataset, the results improved, and the neutralised sentences were more coherent and of a higher quality (Samory et al., 2021). Although we built an automated data augmentation pipeline, it did not improve the per-

	Sequences in order: Input, Gold, Baseline, Fine-tuned, Augmented
1a	women are more likely than men to act in silly ways
1b	some individuals are more likely to act in silly ways
1c	are more likely to appear more likely to act in silly ways
1d	women are equally likely as men to act in silly ways
1e	people are equally likely likely
2a	i cant stand when women commentate on mens sports
2b	i really like when women commentate on mens sports
2c	i can claim when women commentate on mens sports
2d	i cant stand when amateurs commentate on mens sports
2e	i cant stand when people commentate on men
3a	it is ridiculous for a woman to run a locomotive and for a man to darn socks
3b	it is not ridiculous for a woman to run a locomotive and for a man to darn socks
3c	it is ridiculous for a woman to run a locomotive and for a person to darn socks
3d	it is fine for a woman to run a locomotive and for a man to darn socks
3e	it is ridiculous for people to run run a locomotive and a man to darn socks
4a	women are responsible for childhood obesity
4b	parents are responsible for childhood obesity
4c	are responsible for childhood obesity
4d	parents are responsible for childhood obesity
4e	people are responsible for childhood obesity obesity are responsible for childhood people are responsible for people
5a	women must prove they are as good as men
5b	women dont need to prove they are as good as men
5c	women claim to prove they are as good as men
5d	women can prove they are as good as men
5e	people must prove they are as good as anyone else

Table 5: Examples of neutralised sentences for each of the three models for qualitative evaluation.

formance. Finally, we investigated if the model could be evaluated in a purely automated manner. However, we found that a mixture of automatic and human evaluation was required to determine which model performed best. Additional automatic evaluations could be looked into in the future—for example, metrics that consider dissimilarity between the input and predicted sentences. Additionally, we concluded that a combination of metrics was needed to evaluate the model, and future work could investigate how these metrics could be combined. Furthermore, the model could be improved using additional data. It is also worth looking into improving the data augmentation pipeline because much human intervention was required to create the training data. This improved pipeline would then act as the first step in creating the training data. The improved data augmentation pipeline could be used to create initial neutralisations, which could then be manually edited, requiring less edits than manually neutralising the entire dataset.

Finally, this study acts as a first step towards automatically neutralising sexist language and creating a better online environment.

References

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. *Advances in neural information processing systems*, 29.

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

Arun Tejasvi Chaganty, Stephen Mussman, and Percy Liang. 2018. The price of debiasing automatic metrics in natural language evaluation. *arXiv preprint arXiv:1807.02202*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73.

Maeve Duggan. 2017. Online harassment 2017. Accessed: 2021-25-03.

Dylan Grosz and Patricia Conde-Cespedes. 2020. Automatic detection of sexist statements commonly used at the workplace. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 104–115. Springer.

Nizar Habash. 2004. The use of a structural n-gram language model in generation-heavy hybrid machine translation. In *International Conference on Natural Language Generation*, pages 61–69. Springer.

Tatsunori B Hashimoto, Hugh Zhang, and Percy Liang. 2019. Unifying human and statistical evaluation for natural language generation. *arXiv preprint arXiv:1904.02792*.

Akshita Jha and Radhika Mamidi. 2017. When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. In *Proceedings of the second workshop on NLP and computational social science*, pages 7–16.

Jigsaw and Google. Using machine learning to reduce toxicity online. Accessed: 2021-25-03.

Janelle Jones. 2021. 5 facts about the state of the gender pay gap. Accessed: 2021-25-03.

Alexis Krivkovich, Kelsey Robinson, Irina Starikova, Rachel Valentino, and Lareina Yee. 2017. Women in the workplace 2017. Accessed: 2021-25-03.

Irene Langkilde. 2002. An empirical verification of coverage and correctness for a general-purpose sentence generator. In *Proceedings of the international natural language generation conference*, pages 17–24.

Chin-Yew Lin and FJ Och. 2004. Looking for a few good metrics: Rouge and its evaluation. In *Ntcir workshop*.

Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*.

Chandler May, Alex Wang, Shikha Bordia, Samuel R Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. *arXiv preprint arXiv:1903.10561*.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for nlg. *arXiv preprint arXiv:1707.06875*.

Jon Oberlander. 1998. Do the right thing... but expect the unexpected. *Computational Linguistics*, 24(3):501–507.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

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

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1650–1659.

Richard V. Reeves and Ember Smith. 2021. The male college crisis is not just in enrollment, but completion. Accessed: 2021-25-03.

Ehud Reiter and Anja Belz. 2009. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*, 35(4):529–558.

Mattia Samory, Indira Sen, Julian Kohne, Fabian Flöck, and Claudia Wagner. 2021. "call me sexist, but..." : Revisiting sexism detection using psychological scales and adversarial samples. In *ICWSM*.

Ben Schmidt. 2015. Rejecting the gender binary: a vector-space operation. *Ben's Bookworm Blog*.

Donia Scott and Johanna Moore. 2007. An nlg evaluation competition? eight reasons to be cautious. In *Proceedings of the Workshop on Shared Tasks and Comparative Evaluation in Natural Language Generation*, pages 22–23.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the Student Research Workshop, SRW@HLT-NAACL 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 88–93. The Association for Computational Linguistics.

Wikipedia. 2022. Neutral point of view. Accessed: 2021-25-03.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*.