To Prefer or to Choose? Generating Agency and Power Counterfactuals Jointly for Gender Bias Mitigation

Maja Stahl and **Max Spliethöver** and **Henning Wachsmuth**
Leibniz University Hannover, Hannover, Germany
Institute of Artificial Intelligence
{m.stahl, m.spliethoever, h.wachsmuth}@ai.uni-hannover.de

Abstract

Gender bias may emerge from an unequal representation of *agency* and *power*, for example, by portraying women frequently as passive and powerless ("She accepted her future") and men as proactive and powerful ("He chose his future"). When language models learn from respective texts, they may reproduce or even amplify the bias. An effective way to mitigate bias is to generate counterfactual sentences with opposite agency and power to the training. Recent work targeted agency-specific verbs from a lexicon to this end. We argue that this is insufficient, due to the interaction of agency and power and their dependence on context. In this paper, we thus develop a new rewriting model that identifies verbs with the desired agency and power in the context of the given sentence. The verbs' probability is then boosted to encourage the model to rewrite both connotations jointly. According to automatic metrics, our model effectively controls for power while being competitive in agency to the state of the art. In our evaluation, human annotators favored its counterfactuals in terms of both connotations, also deeming its meaning preservation better.

1 Introduction

Gender bias refers to the conscious or unconscious unequal treatment of people because of being male, female, or diverse. In natural language text, it manifests in various ways, including the explicit expression of stereotypes and discrimination as well as implicit prejudicial or generalized representations of genders (Hitti et al., 2019; Doughman et al., 2021). Language models that learn from such text may reproduce or even amplify the bias (Hovy and Spruit, 2016). An effective approach to mitigate this behavior is to reduce bias in the training data (Hitti et al., 2019). In particular, augmenting the data with counterfactuals has been shown to effectively reduce bias in language models (Zmigrod et al., 2019; Lu et al., 2020). Generating counterfactuals that change the depiction of people through the choice of words is the focus of our research.

Several works have analyzed gender bias in the subliminal messages transmitted by the framing of people's actions (Sap et al., 2017; Field et al., 2019; Field and Tsvetkov, 2019; Park et al., 2021). They suggest that the framing of an action influences how the reader perceives the acting person behind. The verb choice can therefore weaken or strengthen the person under consideration (Rashkin et al., 2016; Sap et al., 2017), as in the following example:

1. "She *desires* to get paid." (weakening) vs.
2. "She *demands* to get paid." (strengthening)

To study bias in verb choice, the connotational dimensions of *agency* and *power* as well as their interactions are particularly important (Sap et al., 2017). Agency describes how active a person is portrayed:

3. "X *chooses* their future." (high agency) vs.
4. "X *accepts* their future." (low agency)

Power, on the other hand, describes how much control a person has with respect to a given setting:

5. "X *demands* mercy from their opponent." (high power) vs.
6. "X *begs* their opponent for mercy." (low pow.)

Analyses along these dimensions showed differences between women and men, reflecting gender stereotypes, as detailed in Section 2. For agency-related bias, Ma et al. (2020) created a model that rewrites sentences into a desired agency using the connotation frame lexicon of Sap et al. (2017). We argue that an agency lexicon is not enough to generate counterfactuals, due to the interaction of agency and power and their dependence on context. Especially, power remains untackled so far.

In this paper, we study how to generate counterfactuals for gender bias mitigation by rewriting sentences jointly in terms of both agency and power—while preserving meaning as much as possible. We hypothesize that simply extending an

Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS), pages 39 - 51

agency rewriting model by the power connotation is insufficient to successfully change both connotations of input sentences. Instead, we propose a new model that refines the rewriting process in two ways: First, we determine verbs that are not only similar to the original verb but also have the desired target connotations, by classifying their agency and power within the context of the input sentence. We expect that this results in verbs that allow for a more cohesive sentence rewriting. Second, we boost the generation probability of these verbs for both connotations, encouraging the model to achieve the desired agency and power jointly.

To include verbs indicative of agency and power from diverse contexts, we train the classifiers on sentences from the datasets of Kiesel et al. (2017), Pungas (2017), and Wang et al. (2018). In experiments on the movie summary dataset of Bamman et al. (2013), we then compare our rewriting model against the state-of-the-art for agency (Ma et al., 2020). Concretely, we assess the rewritten sentences in terms of their compliance with target agency, target power, and meaning preservation.

Our automatic pre-evaluation indicates that the new model is competitive in controlling for agency, while outperforming Ma et al. (2020) in terms of power compliance and meaning preservation. In our manual evaluation, human annotators favor our model in terms of all three criteria.

Contributions In summary, our main contributions are:[1]

1. A rewriting model for joint agency and power adaptation on the sentence level.
2. Classifiers for assessing the agency and power of verbs in a given sentence context.
3. Empirical evidence for the importance of joint agency and power control to generate counterfactuals for gender bias mitigation.

Ethical Consideration The methods developed in this paper aim to mitigate gender bias in natural language sentences. As such, we expect primarily positive ethical consequences from the contributions of this paper. However, we point out a significant risk emanating from applying the developed methods: By adjusting the agency and power levels, the meaning of a sentence may likely be changed to some degree. This can have negative implications when facts are distorted. An example of this is misrepresenting a victim as a perpetrator by portraying that person with more agency and/or power. In case our methods are used for modifying language that humans perceive, the methods should thus be used in a semi-automated environment with human supervision. Further ethical implications of this work are discussed in Section 9.

[1]Our code is published at https://github.com/webis-de/NLPANDCSS-22.

2 Related Work

Unequal communication towards social groups, for example in the form of texts, can be the origin of social bias and is one of the main reasons why individuals, their characteristics, and their actions are not perceived correctly. Instead, people's perceptions are often overshadowed and distorted by prejudiced beliefs, resulting in potentially unfair treatment (Steele et al., 2004). Different types of social bias have been studied in NLP research recently (Nangia et al., 2020; Sap et al., 2020; Spliethöver and Wachsmuth, 2020). We focus on one of the most prevailing types, *gender bias*. For comparability with prior work, we use existing datasets in our experiments, limiting them to binary gender instead of considering further social and linguistic gender categories (Cao and Daumé III, 2020).

Previous work analyzed implicit forms of gender bias conveyed through language, often reflected by imbalances in *connotation frames* that capture subjective roles and relations conveyed by a predicate (Rashkin et al., 2016). Connotation frames were introduced by Rashkin et al. (2016), who studied the sentiment and presuppositions of predicates. Sap et al. (2017) extended their notion by explicitly modeling agency and power. The authors created a connotation frame lexicon of common verbs, 2146 of which were manually assigned an agency level, 1737 a power level (positive, equal, or negative). They used the lexicon to compare movie characters, finding that males are generally portrayed with more agency and power. Field et al. (2019) and Field and Tsvetkov (2019) found power imbalances in media articles. For example, female politicians are often portrayed as less powerful than their actual role in society compared to males. Instead of *identifying* gender bias, we focus on *mitigating* it.

Bias mitigation has been addressed at the preprocessing, the training, and the postprocessing level (Feldman and Peake, 2021). One preprocessing approach is to balance gender occurrences in training data. For example, Alhafni et al. (2020) and Sun

et al. (2021) learned on parallel corpora to change the gender of sentences across languages. Park et al. (2018) augmented data with gendered sentences to reduce bias in word embeddings, and Zmigrod et al. (2019) aimed to convert between masculine and feminine inflected sentences without parallel data. At the training level, Dinan et al. (2020) adapted the training process and applied bias controlled training to generative dialogue models to make them generate an equal number of gendered words for both genders considered. Lastly, Bolukbasi et al. (2016), Zhao et al. (2019) and Liang et al. (2020) postprocessed pre-trained word embeddings to remove encoded gender information.

To debias text through the lens of agency connotations, Ma et al. (2020) formalized a new rewriting task called *controllable debiasing* that seeks to correct implicit bias in textual portrayals. Unlike its name (PowerTransformer) suggests, their approach aims to change the *agency* connotation of an input sentence (not power). Using the agency lexicon of Sap et al. (2017), Ma et al. (2020) provide information about agency connotations to the model using self-supervision on a reconstruction task and auxiliary supervision on a paraphrasing task. Inspired by Ghosh et al. (2017), they performed vocabulary boosting at each decoding step based on the agency lexicon to further enhance the agency change.

In this paper, we build on the rewriting model of Ma et al. (2020), but we extend it to jointly control for agency and power. Moreover, we substantially refine the rewriting process by using classifiers to determine which verbs to boost in the given context. Existing classifiers rely on logistic and kernel ridge regression for agency and power, based on decontextualized ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019) embeddings of a verb (Field et al., 2019; Field and Tsvetkov, 2019; Park et al., 2021). In contrast, we use the whole sentence context as input and perform classification, improving the state of the art in our experiments.

3 Approach

This section presents our approach to generating counterfactuals for gender bias mitigation. Based on the contextual classification of verbs, it rewrites sentences jointly in terms of agency and power.

Overview Figure 1 depicts the two parts of the approach: (1) Given a sentence, we identify candidate verbs for its rewritten version. To foster meaning preservation, we filter verbs similar to the original verb. The agency and power of these verbs are then classified in the context of the masked sentence. (2) To rewrite the sentence based on the target connotations, a transformer computes the next-word probability logits. The logits of verbs with target agency and power are boosted to foster connotation change in the output sentence.

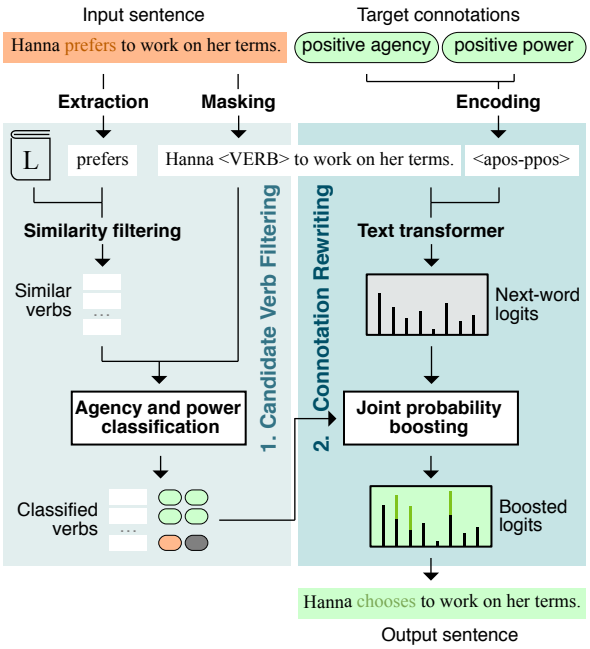


Figure 1: Proposed approach: After masking the verb in the input sentence, all similar verbs from a lexicon L are classified for agency and power in the sentence context. A transformer then rewrites the sentence. At each decoding step, the unnormalized token probabilities (logits) of verbs with target agency and power are increased.

3.1 Candidate Verb Filtering in Context

We seek to find verbs that have a meaning similar to the original verb of a given input sentence s and fit the given target agency and power. As candidates, we consider all verbs from a verb lexicon L.

Similarity Filtering First, we retrieve all verbs from L whose similarity to the original verb in s lies above a threshold γ. Concretely, we employ cosine similarity of the verbs' GloVe representations (Pennington et al., 2014) as a measure.

Agency and Power Classification The next task is to determine the agency and power connotation of all similar verbs. Unlike the lexicon-based connotation filtering of Ma et al. (2020), we classify a verb's agency and power in the context of the masked sentence. In contrast to existing connotation classifiers (Field et al., 2019; Field and

Tsvetkov, 2019), we fine-tune a pre-trained language model based on full sentences. We hypothesize that these changes improve both the identification of the correct agency and power and the resulting cohesiveness of the rewritten sentences.

To emphasize the verb while having the rest of the sentence as context, we separate the verb and the masked sentence with a special token, *[sep]*. The verb is replaced inside the sentence by *[verb]*:

verb [sep] masked_sentence

The resulting sequence is passed to a BERT model (Devlin et al., 2019). As target value, we provide the verb's agency or power connotation as given in the lexicon of Sap et al. (2017). Possible connotation values are *positive*, *equal*, and *negative*.

3.2 Joint Connotation Rewriting

Given a sentence with original agency and power connotations, the task is to rewrite it to express a target agency and power while preserving the meaning as much as possible.

Text Transformer Analog to Ma et al. (2020), we fine-tune a GPT transformer model (Radford et al., 2018) on two tasks: (1) Reconstructing partially masked sentences and (2) paraphrasing sentences. Training for the respective loss functions is done in an alternating manner. For lack of parallel data, the model is trained using self-supervision during reconstruction and auxiliary supervision during paraphrasing. During training, the target agency and power are provided as control tokens, which guide the output connotation during inference. Each control token is composed as follows, where *a* refers to agency and *p* to power, each followed by the respective target value:

<a (pos | equal | neg) – p (pos | equal | neg)>

During reconstruction, the model learns to restore the masked verbs of sentences. Let s be a sentence represented as the sequence of $n \geq 1$ tokens, $s = (t_1, \ldots, t_n)$. The connotations of s are encoded as a control token, t_c. t_c is given to the model as target connotation, along with the masked sentence $\hat{s}$, in which the main verb is replaced by *[verb]*. The target output is the original sentence s. As Ma et al. (2020), we minimize the cross entropy of the target output sentence given the inputs:

$$\mathcal{L}_{recon} = -\frac{1}{n}\sum_{i=1}^{n} \log P(t_i \mid t_1, \ldots, t_{i-1}; \hat{s}; t_c)$$

To enable the model to perform edits that go beyond exchanging verbs, we extend the paraphrasing objective of Ma et al. (2020) whose goal is to achieve coherent, meaningful rewriting. While the verbs in the input sentences are masked as before, t_c now reflects the agency and power connotation of a matching paraphrase $\tilde{s} = (\tilde{t}_1, \ldots, \tilde{t}_m)$, $m \geq 1$. The target output is the paraphrase, $\tilde{s}$. In this way, the control token always represents the connotations of the target output. As with reconstruction, we minimize the cross entropy:

$$\mathcal{L}_{para} = -\frac{1}{m}\sum_{i=1}^{m} \log P(\tilde{t}_i \mid \tilde{t}_1, \ldots, \tilde{t}_{i-1}; \hat{s}; t_c)$$

Joint Probability Boosting At generation time, we boost the probability of verbs with target agency and power to foster the model to change the connotation of a sentence. In this process, the unnormalized probabilities produced by the rewriting model for the next token, called logits $l_i \in \mathbb{R}^{|V|}$ (where V is the vocabulary), are rescaled at each decoding step i to increase the likelihood of generating verbs with the target agency and power. This process is referred to as *boosting*. The boosted logits are then used to compute the next token probabilities:

$$P(t_i|\tilde{t}_1, \ldots, \tilde{t}_{i-1}; s; t_c) \propto softmax(l_i + \beta \cdot A \cdot w)$$

Here, A is a $\mathbb{R}^{V \times 9}$ matrix with a 9-dimensional *{apos-ppos, …, aneg-pneg}* agency-power embedding for each token in the vocabulary V, w is a $\mathbb{R}^9$ one-hot vector encoding of the control token and $\beta \geq 1$ is a scalar hyperparameter representing the boosting strength. Instead of using the connotation frame lexicon as Ma et al. (2020), we encode in A the candidate verbs with target connotations determined by the contextual classification.

4 Data

As part of our experiments, we employ data for two purposes: First, to train the agency and power classifiers, we require sentences that include verbs from the given connotation frame lexicon in a variety of contexts. We therefore combine sentences from three corpora, as detailed below. In our subsequent rewriting experiments, we then use a corpus of movie summaries for the reconstruction objective as well as a parallel paraphrase corpus for the paraphrasing objective. The paraphrase corpus is only used during training, whereas the movie summaries also serve to validate and test rewriting models.

4.1 Data for Agency and Power Classification

We extract all plain-text sentences, which contain any verb indicating agency or power according to the lexicon of Sap et al. (2017),[2] from three existing corpora, covering different contexts and domains:

- Wikipedia biography texts (Wang et al., 2018)
- Plain-text jokes (Pungas, 2017)
- English simple sentences (Kiesel et al., 2017)

As the agency and power labels in the connotation frame lexicon are imbalanced, we undersample the data by removing sentences containing verbs of the majority labels *positive* agency and *positive* power pseudo-randomly. This results in 109,136 sentences labeled for agency and 97,098 for power. A random sample of 20% of the lexicon verbs and the respective sentences are reserved for testing.

4.2 Data for Connotation Rewriting

For our rewriting experiments, we use the movie summary corpus of Bamman et al. (2013). Besides the plain-text plot summaries, the corpus also contains metadata about the movie characters. We use the characters' names and coreferences to perform entity linking. This ensures that each sentence we aim to rewrite contains a character with known agency and power levels. Next, we identify agency and power of each sentence based on its main verb, using the lexicon of Sap et al. (2017). As the main verb, we consider the highest verb in the dependency parse tree of a sentence given by CoreNLP (Manning et al., 2014) that is also the head of a nominal subject dependency with a character mention being the dependent. Finally, we select 25k sentences per gender pseudo-randomly to balance gender occurrences and avoid an underrepresentation of female forms. The total of 50k sentences is then divided into training, validation, and test set using a ratio of 7:2:1.

For the paraphrasing objective, we follow Ma et al. (2020) in taking the parallel corpus by Creutz (2018). For both sentences of each paraphrase pair, we determine the agency and power levels based on the main verb and its associated lexicon entry. The resulting 33,122 pairs are used for training only.

5 Evaluation

This section reports experiments on agency and power classification and on the generation counterfactuals for gender bias mitigation. The main goal is to evaluate our joint agency-power approach to sentence rewriting in light of the state of the art.

5.1 Agency and Power Classification

We tackle the determination of agency and power of a sentence as a three-class tasks each (positive, equal, negative), comparing our contextual classification approach against two baselines:

Approach We trained one BERT model (Base-uncased, 110M parameters) each for agency (*bert-agency*) and power (*bert-power*), using the transformer library of Wolf et al. (2020). We chose to train two separate models, since the correlation between agency and power levels in the lexicon of Sap et al. (2017) is rather low (Kendall's $\tau = 0.30$). We fine-tuned the models in 5-fold cross-validation on the training data from Section 4. To prevent data leakage, we ensured that each verb was included in one fold only. In hyperparameter search, we tested batch sizes from 5 to 35 in increments of 5, learning rates from 10^{-5} to 10^{-9}, and numbers of epochs from 3 to 20. Our final models have been trained using AdamW optimizer (Loshchilov and Hutter, 2019) for 12 epochs with learning rate 10^{-8} and batch size 20, which was the best setting in cross-validation in terms of macro F_1-score.[3]

Baselines We compare our approach to simple majority classifiers (*majority-agency*, *majority-power*) as well as to the state-of-the-art token-level agency and power prediction approach of Field et al. (2019), trained on the given data. We call the latter *log-reg-agency* and *log-reg-power*, since they use logistic regression models for prediction. As input, they employ averaged, and thereby decontextualized, ELMo embeddings of verbs as they appear in training sentences. As ground-truth labels, they also rely on the lexicon of Sap et al. (2017).

Results Table 1 shows the classification results. Our approach achieves the best macro-F_1 scores (0.507 and 0.532 respectively) as well as the highest scores for neutral and negative agency and power. They also reach a more balanced performance across the classes for both target connotations compared to the log-reg baselines.

The confusion matrices in Figure 2 reveal that, if any, our classifiers mostly confuse *positive* or *negative* with *equal* rather than with the opposite class.

[2]The verbs in the sentences are identified using the flair library (Akbik et al., 2019).

[3]All our models were trained on one NVIDIA A100 GPU. Training took about half an hour per epoch for the classifiers.

Model	Positive	Neutral	Negative	Macro
majority-agency	**0.875**	0.000	0.000	0.292
log-reg-agency	0.832	0.146	0.417	0.465
bert-agency (ours)	0.841	**0.252**	**0.430**	**0.507**
majority-power	0.822	0.000	0.000	0.274
log-reg-power	**0.847**	0.272	0.389	0.503
bert-power (ours)	0.805	**0.373**	**0.417**	**0.532**

Table 1: Agency and power classification: Test set macro F_1-scores of our BERT-based classifiers and the baselines, along with the F_1-scores for all three classes.

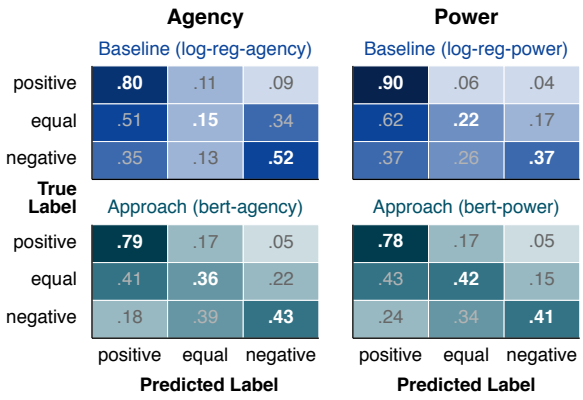


Figure 2: Confusion matrices of the evaluated baseline and our approach for agency and power classification. Our approach avoids strong misclassification notably better, such as classifying negative agency as positive.

In contrast, the log-reg baselines exhibit these more serious errors more often, for example, classifying 37% of the cases with negative power as positive. These results support our hypothesis that sentence context and the pre-trained language understanding of BERT helps differentiate the connotation levels.

5.2 Connotation Rewriting

Next, we test the hypothesis that boosting the candidate verbs found with similarity filtering and contextual classification helps to change both sentence connotations while preserving meaning. To this end, we evaluate the output of rewriting manually.

Approach As Ma et al. (2020), we fine-tuned a pre-trained GPT for 10 epochs on the combined reconstruction and paraphrasing objective. However, we extended the control tokens to also include power (see Section 3). We replicated the training setting of Ma et al. (2020), using AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate of 10^{-5}, a batch size of 4, and top-$p = 0.4$ nucleus sampling (Holtzman et al., 2020) for decoding. To increase the chance of finding suitable candidates, we employed a bigger lexicon than Ma et al. (2020), containing 8,751 verbs.[4] We compared boosting strengths β from 1 to 12 and similarity thresholds γ from 0.2 to 0.5. We found that $\beta = 10$ and $\gamma = 0.5$ effectively control the generation towards the target connotation while minimizing token repetitions.[5]

We also tested a variation of our approach (*Appr. w/o class.*) where we used the lexicon-based connotation filtering of Ma et al. (2020), in combination with our similarity filtering and controlled jointly for agency and power. Here, the boosting strength $\beta = 8$ led to the most promising results.

Baseline We compare our approach to the agency rewriting approach of Ma et al. (2020), trained using the authors' code and settings on the ROC stories corpus (Mostafazadeh et al., 2016) as well as on the paraphrase dataset that we also use to train our model. As previously mentioned, we chose a bigger dataset from a similar domain to train our approach on reconstruction. We hypothesize that this improves the models ability to generate sentences with the desired connotations.

Pre-Evaluation To compare to Ma et al. (2020), we evaluated the approaches first using four of the automatic metrics the authors suggested:[6]

1. *Agency/Power.* Accuracy of changing agency and power, comparing the target connotations to the achieved output connotations according to the lexicon of Sap et al. (2017);[7]
2. *Meaning preservation.* BERTScore F_1 (Zhang et al., 2020), measuring the semantic similarity of input and output sentences;
3. *Fluency.* Perplexity (PPL) of 1000 random output sentences measured using GPT;
4. *Repetition.* The fraction of output sentences containing at least one bigram repetition.

Results Table 2 presents the results of the pre-evaluation. We see that the state-of-the-art baseline performs best in terms of agency change (0.544) and perplexity (134.2). However, its low power accuracy (0.353) reveals that a change in agency

[4] Ma et al. (2020) used the lexicon of Sap et al. (2017) with 2,155 verbs. The list of 8,751 verbs was provided by Ma et al. (2020) for experiments, but did not make it into their model.

[5] Training took about one hour per epoch.

[6] We omitted the fifth measure, uniqueness, as it provides little insight for the scope of this paper.

[7] As the baseline does not control for power separately, we assume target power to equals target agency for its accuracy.

	Agency	Power	Meaning	Fluency	Repetit.
Model	Acc. ↑	Acc. ↑	BScore ↑	PPL ↓	$\text{Rep}_{\geq 2}$ ↓
Ma et al. (2020)	**0.544**	0.353	0.908	**134.2**	0.189
Appr. w/o class.	0.464	**0.495**	**0.931**	161.5	**0.127**
Approach	0.448	0.484	**0.931**	158.2	0.132

Table 2: Automatic pre-evaluation of rewriting quality: Performance of our approach (and its variation without classifiers) on the test set in comparison to the baseline.

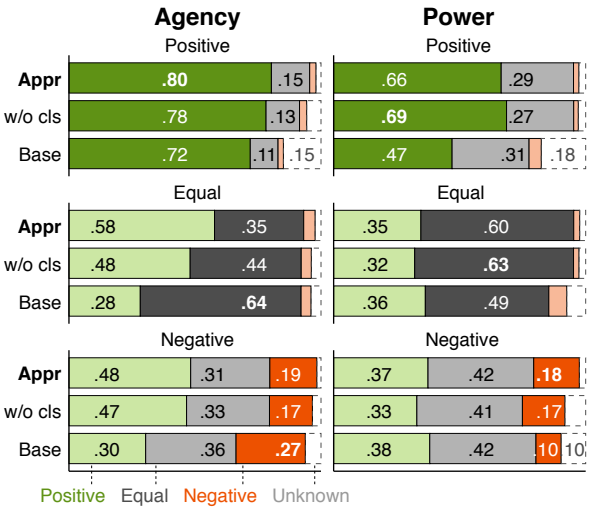


Figure 3: Accuracy of the baseline (*Base*) and both approach variations (*Appr*, Appr *w/o class*) in creating sentences with a specific agency (left) and power (right). The dark-colored bar segment (with white text) in each of the six cases indicates a correct result, the others a wrong one. We beat the baseline in all but two cases.

does not always imply a change in power, stressing the need to control for both connotations. Accordingly, our approach and its variation achieve a much higher power accuracy and similar results on most other metrics. They also preserve the meaning better (BERTScore of 93.1 each) and produce fewer bigram repetitions, resulting in less gibberish sentences that consist of few often repeated tokens.

To understand the models' behavior, Figure 3 shows the agency and power accuracy per target agency and power. Our approach variations perform best on positive agency (.80) and all power levels. Note that this evaluation is unable to assess outputs that do not contain a lexicon verb, including gibberish sentences (shown as *Unknown*).

Main Evaluation The automatic pre-evaluation only roughly approximates quality, especially since it can assess agency and power connotations of lexicon verbs only. We therefore also conducted a user study where six annotators manually evaluated the agency and power change as well as the meaning preservation. All annotators have academic degrees, advanced English skills, and equally represent both genders (no author of this paper).

We selected 450 sentences from the test set randomly, 50 for each of the nine control tokens, that is, for each combination of agency and power connotation. To reduce the workload while remaining able to assess annotation reliability, we divided the sentences into two sets of 225 and let three annotators each evaluate all sentences from one set. We asked all annotators to rank the output sentences by three criteria: *target agency compliance*, *target power compliance*, and *meaning preservation* (annotation guidelines can be found in Appendix A). The average pairwise inter-annotator agreement in terms of Kendall's τ was 0.41 for agency, 0.42 for power, and 0.58 for meaning preservation.

Results Table 3 shows that our *approach* outperforms both other models in terms of all three evaluation criteria. As in the pre-evaluation, it performs similarly to the variation without classifiers on meaning preservation (mean rank 1.69 and 1.73), beating the state of the art of Ma et al. (2020) (2.15). For power and agency, our approach is best with mean rank 1.67 and 1.69 respectively, outperforming Ma et al. (2020) (1.96 and 1.99) again.

The difference to the pre-evaluation in the two latter criteria may be caused by the fact that not all sentences could be evaluated there due to the limitations of the connotation frame lexicon. This speaks for a successful boosting of verbs in general. Another reason lies in the subjectivity of agency and power assessment. While we provided annotators with the same notions of agency and power as previous work, their assessment might still differ from the one encoded in the lexicon.

Ablation Study For further insights, we analyzed the impact of the different parts of our approach on the results. In particular, we compared our full approach to using only the connotation frame lexicon instead of the bigger lexicon for boosting (*No big lex.*), to omitting the similarity filtering (*No sim. filter.*), and to their combination (*Neither*). The results in Table 4 suggest that the connotation frame lexicon would benefit agency and power accuracy. This is expected, since the automatic metrics can only assess those verbs. Omitting similarity filtering seems to worsen the meaning preservation. Our full model scores comparably good for repetition.

Model	Agency Compliance				Power Compliance				Meaning Preservation			
	Rank 1	Rank 2	Rank 3	Mean	Rank 1	Rank 2	Rank 3	Mean	Rank 1	Rank 2	Rank 3	Mean
Ma et al. (2020)	33.5%	36.9%	29.6%	1.96	33.4%	34.4%	32.2%	1.99	23.9%	36.8%	39.3%	2.15
Appr. w/o class.	40.8%	44.4%	14.8%	*1.74	39.8%	43.6%	16.6%	*1.77	44.4%	37.9%	17.8%	*1.73
Approach	**48.7**%	35.0%	16.2%	***1.67**	**46.1**%	39.2%	14.7%	***1.69**	**46.7**%	37.6%	15.7%	***1.69**

Table 3: Manual main evaluation of rewriting quality: Proportion of rewritten sentences with Rank 1, 2, and 3 as well as mean rank per evaluated approach for agency compliance, power compliance, and meaning preservation. Significant gains over Ma et al. (2020) are marked with * (computed using Wilcoxon Signed-Rank Test at $p < .05$).

Model	Agency	Power	Meaning	Fluency	Repetit.
	Acc. ↑	Acc. ↑	B.Sc. ↑	PPL ↓	$\text{Rep}_{\geq 2}$ ↓
No big lex.	**0.452**	0.488	**0.933**	167.0	**0.129**
No sim. filter.	0.407	0.459	0.903	129.8	0.168
Neither	0.445	**0.504**	0.916	**128.4**	0.154
Full model	0.448	0.484	0.931	158.2	0.132

Table 4: Ablation study: Automatic evaluation of rewriting quality for different variations of our approach.

Error Analysis To better understand the differences between the models, we manually inspected some examples from the annotation study. Exemplarily, Table 5 compares three outputs of the models. Matching the automatic results, our approach and its variation generate fewer gibberish sentences with n-gram repetitions than the baseline (see Example 3). Failures in paraphrasing of the latter additionally leads to a reduced meaning preservation (see Example 1). A reason might be the smaller reconstruction dataset, since the paraphrase corpus has the same size. In most cases, the biggest difference between the output sentences is the choice of words (see Example 2), which tends to be best for our approach, according to the annotators.

Lexicon Expansion Lastly, we use our agency and power classifiers to identify potential new verbs for the connotation frame lexicon from the bigger lexicon. Table 6 shows the verbs that are classified to express high or low agency or power with the highest confidence. Most of these partly quite specific verbs match the intuitions of agency and power well. A few classifications may be debatable, though, such as the low agency for "thrive".

6 Conclusion

In this work, we have studied how to rewrite sentences to adjust the agency and power of their subjects jointly. To this end, we have developed a new candidate verb identification method that fosters a meaning-preserving adaptation of both connotations in transformer-based generation. By employing classifiers for agency and power, our rewriting approach can handle any given verb in the current sentence context, unlike previous approaches.

Our experiments have stressed the importance of addressing agency and power jointly. In automatic evaluation, the proposed approach has turned out competitive in agency to the previous state of the art, while effectively controlling for power for the first time. In manual evaluation, human annotators favored the sentences rewritten by our approach in terms of all relevant dimensions: target agency, target power, and meaning preservation.

We thus conclude that our approach contributes towards the generation of counterfactuals that can be used for gender bias mitigation, as shown in previous work. Yet, the results leave room for improvements regarding both connotations, which should be addressed in future work. For a refined evaluation, more extensive agency and power lexicons may be needed. To facilitate the lexicon creation process, the classifiers may be used to suggest labels prior to human post-checking.

7 Acknowledgments

We thank the anonymous reviewers, particularly for their valuable comments on potential ethical concerns. Moreover, we thank Malena Brinkmann, Niklas Hahn, Eileen Köllmann, Rica Nieland, Jonas Rieskamp and Niclas Stahl for their feedback. This work has been partially funded by the German Research Foundation (DFG) within the project ArgSchool, project number 453073654.

8 Limitations

The study of counterfactual generation carried out in this paper has a number of limitations, the most important of which we discuss in the following.

First, for a clear paper focus, we decided to leave out any experiments on the impact of the generated counterfactuals on actual gender bias mitigation. While previous research has already shown that the

#	Input Sentence	Target	Model	Output Sentence
1	Johnny, enraged with the way Joe and Jake treat Ginny, *[verb]* them.	apos-ppos	Ma et al. (2020)	pos hurt ginny, hurting them.
			Appr. w/o cls.	johnny, enraged with the way joe and jake treat ginny, attacks them.
			Approach	johnny, enraged with the way joe and jake treat ginny, defend them.
2	After some time Tristan is *[verb]* to flee to Cornwall.	aneg-pequal	Ma et al. (2020)	tristan is forced to flee to cornwall after a period of time.
			Appr. w/o cls.	tristan is seen to flee to cornwall after a period of time.
			Approach	tristan is sensed to flee to cornwall after a period of time.
3	Enraged, Roopa *[verb]* to commit suicide by jumping into a waterfall as she finds it better to kill herself rather than be a mistress.	aequal-ppos	Ma et al. (2020)	pos decided to commit suicide by jumping into a waterfall as she becomes a mistress better to kill herself better to kill herself better to kill herself better to kill herself better to kill herself better to kill herself better to kill kill kill kill kill kill kill kill kill kill kill ...
			Appr. w/o cls.	enraged, roopa resolves to commit suicide by jumping into a waterfall as she finds it better to kill herself rather than be a mistress
			Approach	enraged, roopa seeks to commit suicide by jumping into a waterfall as she finds it better to kill herself rather than be a mistress.

Table 5: Output sentences generated by the three rewriting models for the given masked input sentences and target connotations: (1) The baseline (Ma et al., 2020) attempts to paraphrase. Both approach variations generate similar output; (2) our approach matches the target connotation; (3) the baseline starts promising but ends with gibberish.

	Agency-related Verbs			Power-related Verbs		
high	inflate	gun	instate	ionize	ambush	damn
	bridge	reset	distance	recycle	augment	indict
	extort	maim	reanimate	fracture	reset	auction
low	bloom	reside	succumb	repent	revere	venerate
	thrive	average	yearn	profess	elate	mediate
	aspire	crave	slumber	rejoice	yearn	heed

Table 6: New candidate verbs for the connotation frame lexicon, selected based on the classification and confidence level of our agency and power classifiers.

intended use of counterfactuals helps in this regard, we therefore can ultimately not make assertions on the practical benefit of our method compared to others. Future work should investigate upon the use of our method in downstream tasks.

Furthermore, the unequal portrayal of people in terms of their agency and power represents only one of different ways of how gender bias manifests in language. As a matter of fact, even if our method was perfectly effective, it would not suffice alone to mitigate gender bias to the full extent. Moreover, despite the evidence we found that the proposed method improves over the state of the art in terms of agency and power rewriting, its effectiveness still shows notable room for improvement. It is noteworthy, though, that our experiments suggest that the method rarely modifies agency and power in the opposite direction, implying that additional harm caused by the method is unlikely.

Finally, the data used in our experiments restricts the generalizability of our results to some extent. In particular, we analyzed the effectiveness of our method on English movie summaries only. Other genres as well as other languages may lead to different behavior, although we do not see an immediate reason why it should not work there. As far as the availablity of data will permit, we plan to do further experiments in other settings in the future.

9 Ethical Statement

The intended use of the methods developed in this paper is to mitigate gender bias in natural language sentences. The goal of applying these methods is to obtain linguistic data that allows for a more equal representation of genders (e.g., for the training of embedding models). As such, we predominantly expect positive ethical consequences of the contributions of this paper. However, we see two noteworthy risks that emanate from an availability of the developed methods:

First, due to the methods' non-perfect effectiveness and to the general complexity of natural language, the bias mitigation may come with possibly unintended changes of meaning of the sentences being rewritten. This may lead to a misrepresentation of genders or specific representatives of the genders. The effects of applying the methods should therefore be observed carefully. Where possible, it should ideally be in a semi-automatic setting with human post-checking.

Second, as with any other text generation technology, the methods may be misused for an application they are not meant for, for example, to picture an individual or a group of people in a mislead-

ing way. We cannot prevent such usage, but the still limited effectiveness of the methods makes a purposeful deceptive usage in our view impractical.

Aside from the risks, we would like to state explicitly again that the consideration of gender as a binary dimension, as done in this paper, is a simplification of reality. The only reason why we restrict our view exclusively to men and women is the lack of data for studying tasks as the given one more properly with respect to gender diversity.

Finally, we point out that no personal information has been collected from any participant of our annotation study; there is no way to match the created annotations to their identities. The participants came from the surroundings of authors of this paper. Participation was not paid for, but more done in terms of a friendly turn. It was fully voluntary.

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. Flair: An easy-to-use framework for state-of-the-art nlp. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.

Bashar Alhafni, Nizar Habash, and Houda Bouamor. 2020. Gender-aware reinflection using linguistically enhanced neural models. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 139–150, Barcelona, Spain (Online). Association for Computational Linguistics.

David Bamman, Brendan O'Connor, and Noah A. Smith. 2013. Learning latent personas of film characters. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 352–361, Sofia, Bulgaria. 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*, volume 29, pages 4349–4357, Barcelona Spain. Curran Associates, Inc.

Yang Trista Cao and Hal Daumé III. 2020. Toward Gender-Inclusive Coreference Resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4568–4595, Online. Association for Computational Linguistics.

Mathias Creutz. 2018. Open subtitles paraphrase corpus for six languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2020. Queens are powerful too: Mitigating gender bias in dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8173–8188, Online. Association for Computational Linguistics.

Jad Doughman, Wael Khreich, Maya El Gharib, Maha Wiss, and Zahraa Berjawi. 2021. Gender bias in text: Origin, taxonomy, and implications. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 34–44, Online. Association for Computational Linguistics.

Tal Feldman and Ashley Peake. 2021. End-To-End Bias Mitigation: Removing Gender Bias in Deep Learning. *arXiv:2104.02532 [cs]*. http://arxiv.org/abs/2104.02532.

Anjalie Field, Gayatri Bhat, and Yulia Tsvetkov. 2019. Contextual affective analysis: A case study of people portrayals in online #metoo stories. *Proceedings of the International AAAI Conference on Web and Social Media*, 13(01):158–169.

Anjalie Field and Yulia Tsvetkov. 2019. Entity-centric contextual affective analysis. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2550–2560, Florence, Italy. Association for Computational Linguistics.

Sayan Ghosh, Mathieu Chollet, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer. 2017. Affect-LM: A neural language model for customizable affective text generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 634–642, Vancouver, Canada. Association for Computational Linguistics.

Yasmeen Hitti, Eunbee Jang, Ines Moreno, and Carolyne Pelletier. 2019. Proposed taxonomy for gender bias in text; a filtering methodology for the gender generalization subtype. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 8–17, Florence, Italy. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text

degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

Johannes Kiesel, Benno Stein, and Stefan Lucks. 2017. A large-scale analysis of the mnemonic password advice. In *24th Annual Network and Distributed System Security Symposium, NDSS 2017, San Diego, California, USA, February 26 - March 1, 2017*. The Internet Society.

Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Towards Debiasing Sentence Representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5502–5515, Online. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2020. *Gender Bias in Neural Natural Language Processing*, pages 189–202. Springer International Publishing, Cham.

Xinyao Ma, Maarten Sap, Hannah Rashkin, and Yejin Choi. 2020. PowerTransformer: Unsupervised controllable revision for biased language correction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7426–7441, Online. Association for Computational Linguistics.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Chan Young Park, Xinru Yan, Anjalie Field, and Yulia Tsvetkov. 2021. Multilingual contextual affective analysis of LGBT people portrayals in wikipedia. In *Proceedings of the Fifteenth International AAAI Conference on Web and Social Media, ICWSM 2021, held virtually, June 7-10, 2021*, pages 479–490. AAAI Press.

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, Brussels, Belgium. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics.

Taivo Pungas. 2017. A dataset of English plaintext jokes.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training.

Hannah Rashkin, Sameer Singh, and Yejin Choi. 2016. Connotation frames: A data-driven investigation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 311–321, Berlin, Germany. Association for Computational Linguistics.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Maarten Sap, Marcella Cindy Prasettio, Ari Holtzman, Hannah Rashkin, and Yejin Choi. 2017. Connotation frames of power and agency in modern films.

In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2329–2334, Copenhagen, Denmark. Association for Computational Linguistics.

Maximilian Spliethöver and Henning Wachsmuth. 2020. Argument from old man's view: Assessing social bias in argumentation. In *Proceedings of the 7th Workshop on Argument Mining*, pages 76–87, Online. Association for Computational Linguistics.

Jennifer Steele, Y. Susan Choi, and Nalini Ambady. 2004. Stereotyping, Prejudice, and Discrimination. In Theresa A. Thorkildsen and Herbert J. Walberg, editors, *Nurturing Morality*, Issues in Children's and Families' Lives, pages 77–97. Springer, Boston, MA.

Tony Sun, Kellie Webster, Apurva Shah, William Yang Wang, and Melvin Johnson. 2021. They, them, theirs: Rewriting with gender-neutral english. *CoRR*, abs/2102.06788.

Qingyun Wang, Xiaoman Pan, Lifu Huang, Boliang Zhang, Zhiying Jiang, Heng Ji, and Kevin Knight. 2018. Describing a knowledge base. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 10–21, Tilburg University, The Netherlands. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender Bias in Contextualized Word Embeddings. 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 629–634, Minneapolis, Minnesota. Association for Computational Linguistics.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy. Association for Computational Linguistics.

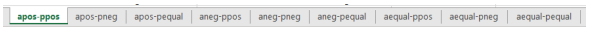


Figure 4: Sheets in the annotation file. You can switch between sheets at the bottom of the window.

A Human Annotation Guidelines

A.1 Introduction

This document contains instructions for the annotation study that is being conducted in the course of our research. Additionally, the concepts relevant to performing the tasks are explained. The goal of this annotation study is to rank the output sentences generated by three different models.

These models were developed with the goal of rewriting input sentences so that they subsequently express the target agency and target power, which will be explained in the following sections. At the same time, the meaning of the input sentences should remain the same as far as possible.

A.2 What is Agency?

The agency level of a sentence describes how **active**, **decisive** or **energetic** the main person of the sentence is portrayed as. High agency stands for activity, while low agency stands for passivity. In the example in Table 7, the name "X" and neutral pronouns were chosen to avoid triggering gender bias.

A.3 What is Power?

The power level of a sentence describes how **powerful**, **strong** or **influential** the main person of the sentence is portrayed as. A distinction is made between whether the main person has power over the theme (high power) or whether the theme has power over the main person (low power).

As can be seen from the example "He begged his opponent for mercy." (see Table 8) a sentence can express different levels of agency and power simultaneously, but this need not be the case.

A.4 Instructions

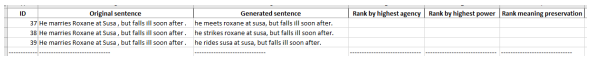
Your task will be to **rank** the agency, power and meaning preservation of three **generated sentences** per one original sentence. You will receive an Excel file containing the sentences that should be annotated. This file will contain nine sheets with different sentences (see Figure 4)

Example	Agency	Explanation
X *chose* their future.	high	X is actively choosing and taking charge of their future.
X *begged* their opponent for mercy.	high	X is actively trying to invoke mercy.
X *demanded* mercy from their opponent	high	X is actively trying to invoke mercy.
X *accepted* their future.	low	X passively agrees to what is happening.
X *survived* the crash.	low	X is portrayed as not having active influence on their survival.
X used to *fear* dogs.	low	X's fear was not actively influenceable.

Table 7: Agency examples.

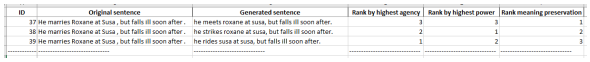
Example	Power	Explanation
X *demanded* mercy from their opponent	high	X tells the opponent what to do and has therefore power over them.
X *chose* their future.	high	X has power over their future because they shape the future themself.
X *hugs* their father.	high	X is portrayed as influencing the interaction with their father.
X *begged* their opponent for mercy.	low	The opponent is portrayed as having power over them.
X *admitted* their mistake.	low	The mistake influences X's actions.
X used to *fear* dogs.	low	Dogs have power over X instead of the other way around.

Table 8: Power examples.



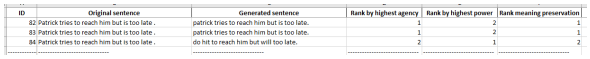
ID	Original sentence	Generated sentence	Rank by highest agency	Rank by highest power	Rank meaning preservation
37	He marries Roxane at Susa , but falls ill soon after .	he meets roxane at susa, but falls ill soon after.			
38	He marries Roxane at Susa , but falls ill soon after .	he strikes roxane at susa, but falls ill soon after.			
39	He marries Roxane at Susa , but falls ill soon after .	he rides susa at susa, but falls ill soon after.			

Figure 5: Header and first example in sheet "apos-ppos".



ID	Original sentence	Generated sentence	Rank by highest agency	Rank by highest power	Rank meaning preservation
37	He marries Roxane at Susa , but falls ill soon after .	he meets roxane at susa, but falls ill soon after.	3	3	1
38	He marries Roxane at Susa , but falls ill soon after .	he strikes roxane at susa, but falls ill soon after.	2	1	2
39	He marries Roxane at Susa , but falls ill soon after .	he rides susa at susa, but falls ill soon after.	1	2	3

Figure 6: Possible annotations for the first example.



ID	Original sentence	Generated sentence	Rank by highest agency	Rank by highest power	Rank meaning preservation
82	Patrick tries to reach him but is too late .	patrick tries to reach him but is too late.	1	2	1
83	Patrick tries to reach him but is too late .	patrick tries to reach him but is too late.	1	2	1
84	Patrick tries to reach him but is too late .	do hit to reach him but will too late.	2	1	2

Figure 7: Possible annotations for another example, in which two generated sentences are equal.

All nine sheets should be filled in. Figure 5 shows an example on the first sheet "apos-ppos" of how the header and the first example might look like.

The ID column can be ignored. The first relevant column contains the original sentence, which should be used as reference to rate the meaning preservation. For each group of three generated sentences, the original sentence will be the same. Next, the three generated sentences are displayed. Those should be read carefully to then **rank the agency, the power and the meaning preservation from 1-3 comparing the generated sentences with each other**. In this example, the sentence with the highest agency should get rank 1, the sentence with the next highest agency rank 2 and the remaining one rank 3. Same goes for power and meaning preservation (see Figure 6).

On each sheet, the agency and power assessment tasks are slightly different. The possible variations are:

1. Rank by **highest** agency / power
2. Rank by **most neutral** agency / power
3. Rank by **lowest** agency / power

As the instructions suggest, for "the most neutral" the sentence with the most neutral agency/power should get rank 1. The same goes for "lowest", where the sentence with the lowest agency/power should be ranked 1.

In case you feel like two or more sentences should have the **same ranking in one or more category**, because the agency, power and/or meaning preservation is the same, feel free to give them the same score. In the following example, two models created the same sentence, which leads to the same annotation for them. But it could also be different sentences for which you feel like the agency, power or meaning preservation are equal (see Figure 7).