Exploring the Efficacy of Automatically Generated Counterfactuals for Sentiment Analysis

Linyi Yang [1,2,3,4], **Jiazheng Li** [2], **Pádraig Cunningham** [2], **Yue Zhang** [3,4]

Barry Smyth [1,2], **Ruihai Dong** [1,2]

[1] The Insight Centre for Data Analytics, University College Dublin
[2] School of Computer Science, University College Dublin
[3] School of Engineering, Westlake University
[4] Institute of Advanced Technology, Westlake Institute for Advanced Study
{linyi.yang, ruihai.dong, barry.smyth}@insight-centre.org
{padraig.cunningham}@ucd.ie
{jiazheng.li}@ucdconnect.ie
{yue.zhang}@westlake.edu.cn

Abstract

While state-of-the-art NLP models have been achieving the excellent performance of a wide range of tasks in recent years, important questions are being raised about their robustness and their underlying sensitivity to systematic biases that may exist in their training and test data. Such issues come to be manifest in performance problems when faced with out-of-distribution data in the field. One recent solution has been to use counterfactually augmented datasets in order to reduce any reliance on spurious patterns that may exist in the original data. Producing high-quality augmented data can be costly and time-consuming as it usually needs to involve human feedback and crowdsourcing efforts. In this work, we propose an alternative by describing and evaluating an approach to automatically generating counterfactual data for the purpose of data augmentation and explanation. A comprehensive evaluation on several different datasets and using a variety of state-of-the-art benchmarks demonstrate how our approach can achieve significant improvements in model performance when compared to models training on the original data and even when compared to models trained with the benefit of human-generated augmented data.

1 Introduction

Deep neural models have recently made remarkable advances on sentiment analysis (Devlin et al., 2018; Liu et al., 2019; Yang et al., 2019; Xie et al., 2020). However, their implementation in practical applications still encounters significant challenges. Of particular concern, these models tend to learn intended behavior that is often associated with spurious patterns (artifacts) (Jo and Bengio, 2017; Slack et al., 2020a). As an example, in the sentence *"Nolan's films always shock people, thanks to his superb directing skills"*, the most influential word for the prediction of a positive sentiment should be *"superb"* instead of *"Nolan"* or *"film"*. The issue of spurious patterns also partially affects the out-of-domain (OOD) generalization of the models trained on independent, identical distribution (IID) data, leading to performance decay under distribution shift (Quionero-Candela et al., 2009; Sugiyama and Kawanabe, 2012; Ovadia et al., 2019).

Researchers have recently found that such concerns about model performance decay and social bias in NLP come about out-of-domain because of a sensitivity to semantically spurious signals (Gardner et al., 2020), and recent studies have uncovered a problematic tendency for gender bias in sentiment analysis (Zmigrod et al., 2019; Maudslay et al., 2019; Lu et al., 2020). To this end, one of the possible solutions is data augmentation with counterfactual examples (Kaushik et al., 2020) to ensure that models learn real causal associations between the input text and labels. For example, a sentiment-flipped counterfactual of last example could be *"Nolan's movies always **bore** people, thanks to his **poor** directorial skills."*. When added to the original set of training data, such kinds of counterfactually augmented data (CAD) have shown their benefits on learning real causal associations and improving the model robustness in recent studies (Kaushik et al., 2020, 2021; Wang and Culotta, 2021). Unlike gradient-based adversarial examples (Wang and Wan, 2019; Zhang et al., 2019; Zang et al., 2020), which cannot provide a clear boundary between positive and negative instances to humans, counterfactuals could provide *"human-like"* logic to show a modification to the

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics
and the 11th International Joint Conference on Natural Language Processing, pages 306–316
August 1–6, 2021. ©2021 Association for Computational Linguistics

input that makes a difference to the output classification (Byrne, 2019).

Recent attempts for generating counterfactual examples (also known as minimal pairs) rely on human-in-the-loop systems. Kaushik et al. (2020) proposed a human-in-the-loop method to generate CAD by employing human annotators to generate sentiment-flipped reviews. The human labeler is asked to make minimal and faithful edits to produce counterfactual reviews. Similarly, Srivastava et al. (2020) presented a framework to leverage strong prior (human) knowledge to understand the possible distribution shifts for a specific machine learning task; they use human commonsense reasoning as a source of information to build a more robust model against spurious patterns. Although useful for reducing sensitivity to spurious correlations, collecting enough high-quality human annotations is costly and time-consuming.

The theory behind the ability of CAD to improve model robustness in sentiment analysis is discussed by Kaushik et al. (2021), where researchers present a theoretical characterization of the impact of noise in causal and non-causal features on model generalization. However, methods for automatically generating CAD have received less attention. The only existing approach (Wang and Culotta, 2021) has been tested on the logistic regression model only, despite the fact that recent state-of-the-art methods for sentiment classification are driven by neural models. Also, their automatically generated CAD cannot produce competitive performance compared to human-generated CAD. We believe that their method does not sufficiently leverage the power of pre-trained language models and fails to generate fluent and effective CAD. In addition, the relationships between out-of-domain generalization and sensitivity to spurious patterns were not explicitly investigated by Wang and Culotta (2021).

To address these issues, we use four benchmark datasets (IMDB movie reviews as hold-out test while Amazon, Yelp, and Twitter datasets for out-of-domain generalization test) to further explore the efficacy of CAD for sentiment analysis. First, we conduct a systematic comparison of several different state-of-the-art models (Wang and Culotta, 2021). This reveals how large Transformer-based models (Vaswani et al., 2017) with larger parameter sizes may improve the resilience of machine learning models. Specifically, we have found that for increasing parameter spaces, CAD's performance benefit tends to decrease, regardless of whether CAD is controlled manually or automatically. Second, we introduce a novel masked language model for helping improve the fluency and grammar correctness of the generated CAD. Third, we add a fine-tuned model as a discriminator for automatically evaluating the edit-distance, using data generated with minimal and fluent edits (same requirements for human annotators in Kaushik et al. (2020)) to ensure the quality of generated counterfactuals. Experimental results show that it leads to significant prediction benefits using both hold-out tests and generalization tests.

To the best of our knowledge, we are the first to automatically generate counterfactuals for use as augmented data to improve the robustness of neural classifiers, which can outperform existing, state-of-the-art, human-in-the-loop approaches. We will release our code and datasets on GitHub [1].

2 Related Work

This work mainly touches on three important areas: approaches to evaluation that go beyond traditional accuracy measures (Bender and Koller, 2020; Warstadt et al., 2020), the importance of counterfactuals in eXplainable AI (XAI) (Byrne, 2019; Keane and Smyth, 2020), and out-of-domain generalization in sentiment analysis (Kim and Hovy, 2004; Zhang et al., 2018; Zhang and Zhang, 2019).

There has been an increasing interest in the role of **Robustness Causal Thinking** in ML, often by leveraging human feedback. Recently, some of the standard benchmark datasets have been challenged (Gardner et al., 2020; Ribeiro et al., 2020), in which the model performance is significantly lower on contrast sets than on original test sets; a difference of up to 25% in some cases. Researchers propose counterfactual data augmentation approaches for building robust models (Maudslay et al., 2019; Zmigrod et al., 2019; Lu et al., 2020), and find that spurious correlations threaten the model's validity and reliability. In an attempt to address this problem, Kaushik et al. (2020) explore opportunities for developing human-in-the-loop systems by using crowd-sourcing to generate counterfactual data from original data, for data augmentation. Teney et al. (2020) shows the continuous effectiveness of CAD in computer vision (CV) and NLP.

The idea of generating **Counterfactuals in XAI**

[1] https://github.com/lijiazheng99/Counterfactuals-for-Sentiment-Analysis

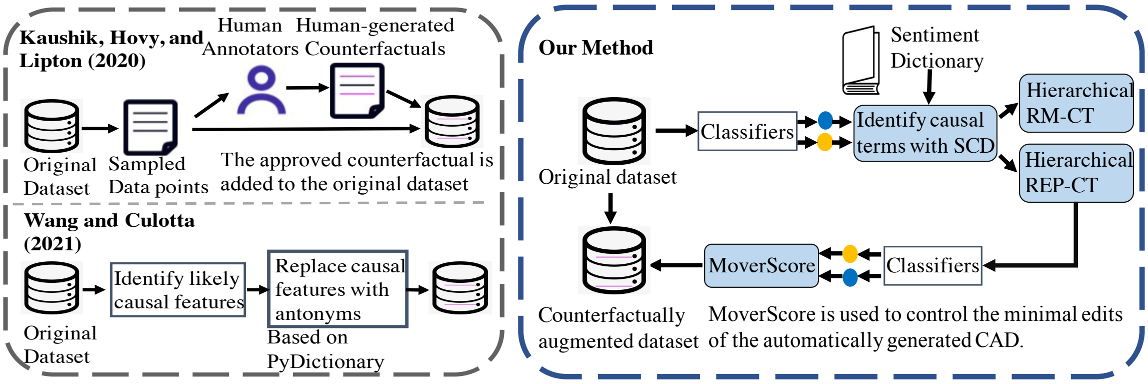
Kaushik, Hovy, and Lipton (2020)
Human Annotators
Human-generated Counterfactuals
Original Dataset
Sampled Data points
The approved counterfactual is added to the original dataset
Wang and Culotta (2021)
Original Dataset
Identify likely causal features
Replace causal features with antonyms
Based on PyDictionary
Our Method
Sentiment Dictionary
Original dataset
Classifiers
Identify causal terms with SCD
Hierarchical RM-CT
Hierarchical REP-CT
Counterfactually augmented dataset
MoverScore
Classifiers
MoverScore is used to control the minimal edits of the automatically generated CAD.

Figure 1: Overview of previous CAD methods are shown on the left side, while the pipeline of our method is shown on the right. Hierarchical RM-CT (removing the casual terms) and Hierarchical REP-CT (replacing the casual terms) are our methods for automatically generating CAD, respectively. SCD denotes *sampling and sensitivity of contextual decomposition*. Sentiment Dictionary refers to the opinion lexicon published by (Hu and Liu, 2004).

also shares important conceptual features with our work. Since human counterfactual explanations are minimal in the sense that they select a few relevant causes (Byrne, 2019; Keane and Smyth, 2020) as is the requirement of minimal edits in our generation process. This has been explored more in the field of CV (Goyal et al., 2019; Kenny and Keane, 2021), but investigated less in NLP. Recent work (Jacovi and Goldberg, 2020) highlight explanations of a given causal format, and Yang et al. (2020a) generate counterfactuals for explaining the prediction of financial text classification. We propose a similar but different research question, that is, whether the automatically generated counterfactual can be used for data augmentation to build more robust models, which has not been considered by the previous methods in XAI (Pedreschi et al., 2019; Slack et al., 2020b; Yang et al., 2020b; Ding et al., 2020).

In the case of **Sentiment Analysis**, most of the previous works report experiments using a hold-out test on the IID dataset (Liu, 2012; Yang et al., 2016; Johnson and Zhang, 2017). The current state-of-the-art methods make use of large pre-trained language models (e.g., BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019) and SMART-RoBERTa (Jiang et al., 2020)) for calculating input represntations. It has been shown that these methods can suffer from spurious patterns (Kaushik et al., 2020; Wang and Culotta, 2021). Very recently, Wang and Culotta (2021) provide a starting point for exploring the efficacy of automatically generated CAD for sentiment analysis, but it is still based on IID hold-out tests only. However, spurious patterns in the training and test sets could be tightly coupled, which may limit the possibility of observing their attendant accuracy issues using a hold-out test methodology. For this reason, we designed an indirect method for evaluating the robustness of models, by comparing the performance of models trained on original and augmented data using out-of-domain data. The prediction benefit for out-of-domain data should provide some evidence about whether a model's sensitivity to spurious patterns has been successfully mitigated. The resulting counterfactuals can be used for data augmentation and can also provide contrastive explanations for classifiers, and important and desirable consideration for the recent move towards more XAI (Ribeiro et al., 2016; Lundberg and Lee, 2017; Lipton, 2018; Pedreschi et al., 2019; Slack et al., 2020b).

3 Detailed Implementation

We propose a new approach for automatically generating counterfactuals to enhance the robustness of sentiment analysis models by inverting the sentiment of causally important terms according to Algorithm 1 and based on the following stages:

1. The identification of genuine causal terms using self-supervised contextual decomposition (Section 3.1).
2. Generating counterfactual samples by (a) RM-CT (removing causal terms) and (b) REP-CT (replacing the causal terms) (Section 3.2).
3. Selecting the human-like counterfactuals using MoverScore. (Zhao et al., 2019) (Section 3.3).

The end result will be a set of counterfactuals that can be used to augment an existing dataset.

3.1 Identifying Causal Terms

To identify causally important terms, we propose a hierarchical method, based on the *sampling and sensitivity of contextual decomposition* technique from Jin et al. (2019), by incrementally removing words from a sentence in order to evaluate the model's sensitivity to these words. Significant changes in model outputs suggest the removal of important terms. For example, removing the word *"best"* from *"The movie is the best that I have ever seen."*, is likely to alter a model's sentiment prediction more than the removal of other words from the sentence; thus *"best"* is an important word with respect to this sentence's sentiment. In a similar way, phrases beginning with negative pronouns will likely be important; for instance, *"not satisfy you"* is important in *"This movie could not satisfy you"*.

Given a word (or phrase starting with negative limitations) w in the sentence s, the importance of w can be calculated as in Equation 1 where $\mathbf{s}_{\text{-}\beta}\backslash\mathbf{p}$ denotes the sentence that resulting after masking out a single word (or a negative phrase as above). We use $l\left(\mathbf{s}_{\text{-}\beta}\backslash\mathbf{p};\widehat{\mathbf{s}}\right)$ to represent the model prediction after replacing the masked-out context, while $\widehat{\mathbf{s}_{\beta}}$ is a input sequence sampled from the input $\mathbf{s}$. $\backslash\mathbf{p}$ indicates the operation of masking out the phrase p in a input document $\mathcal{D}$ from the training set. The specific candidate causal terms found by this masking operation vary for different prediction models.

$$\phi(\mathbf{w},\widehat{\mathbf{s}}) = \mathbb{E}_{\mathbf{s}_{\beta}}\left[\frac{l\left(\mathbf{s}_{\text{-}\beta};\widehat{\mathbf{s}_{\beta}}\right) - l\left(\mathbf{s}_{\text{-}\beta}\backslash\mathbf{p};\widehat{\mathbf{s}_{\beta}}\right)}{l\left(\mathbf{s}_{\text{-}\beta};\widehat{\mathbf{s}_{\beta}}\right)}\right] \quad (1)$$

3.2 Generating Human-like Counterfactuals

This approach and the scoring function in Equation 1 is used in Algorithm 1 in two ways, to generate two types of plausible counterfactuals. First, it is used to identify words to *remove* from a sentence to produce a plausible counterfactual. This is referred to as RM-CT and is performed by lines 3–5 in Algorithm 1; for a sentence $S^{(i)}$, it's correctly labeled sentiment words are identified (line 3), and sorted based on Equation 1 (line 4) with classifier C, and the most important of these words is removed from $S^{(i)}$ to produce $S^{(i)}_{rm}$ (line 5).

Algorithm 1 Generating plausible counterfactual instances.

Input: Test document $\mathcal{D}^{(n)}$= $\{P_1, P_2, ..., P_n\}$, with corresponding ground-truth labels **Y**, pre-trained Mask Language Model **MLM**, fine-tuned transformer classifier **C**, Positive Word Dictionaries **POS**, Negative Word Dictionaries **NEG**. (**pos** and **neg** are predicates for positive and negative labels)
Output: Plausible counterfactual $D^{(k)}_{cf}$= $\{D^{(k)}_{rep}, D^{(k)}_{rm}\}$

```
 1: for $P_k$ in $D^{(n)}$ do
 2:     for $S^{(i)}, Y_i$ in $P_k$ do
 3:         $\widehat{S}^{(i)} \leftarrow \{w \in S^{(i)} \mid (w \in POS \wedge Y_i = pos)$
                    $\vee\ (w \in NEG \wedge Y_i = neg)\}$
 4:         $S^{(i)}_{sorted} \leftarrow sort\big(\widehat{S}^{(i)}, key = \phi(w, \widehat{S}^{(i)})\big)(eq.1)$
 5:         $S^{(i)}_{rm} \leftarrow S^{(i)}_{sorted}[1:]$
 6:         $S^{(i)}_{rep} \leftarrow S^{(i)}_{sorted}$
 7:         for $w \in S^{(i)}_{rep}$ do
 8:             $W_p \leftarrow MLM\big(S^{(i)}_{mask(w)}, S^{(i)}_{rep}\big)$
 9:             $W_c \leftarrow \{w \in W_p \mid (w \in POS \wedge Y_i! =$
    $pos) \vee (w \in NEG \wedge Y_i! = neg)\}$
10:             $S^{(i)}_{rep}(w) \leftarrow sort\big(W_c, key = \phi(w, W_c)\big)[0]$
11:         end for
12:         $P^{(k)}_{rm} \leftarrow P^{(k)}_{rm} + S^{(i)}_{rm}$
13:         $P^{(k)}_{rep} \leftarrow P^{(k)}_{rep} + S^{(i)}_{rep}$
14:     end for
15:     $D^{(n)}_{rm} \leftarrow D^{(n)}_{rm} + P^{(k)}_{rm}$
16:     $D^{(n)}_{rep} \leftarrow D^{(n)}_{rep} + P^{(k)}_{rep}$
17: end for
18: return $D^{(n)}_{rm}, D^{(n)}_{rep}$
```

Second, the REP-CT technique instead *replaces* each causally important sentiment word in $S^{(i)}$ with an alternative word that has an opposing sentiment polarity (lines 6-11 in Algorithm 1). To do this the words in $S^{(i)}$ are each considered for replacement in order of their importance (lines 6 & 7) to create a new sentence $S^{(i)}_{rep}$. For each word w we use a masked language model (MLM) to generate a set of plausible replacements, W_p (line 8), and a subset of these, W_c, as replacement candidates if their sentiment is different from the sentiment of $S^{(i)}$, which is given by Y_i (line 9). Here we are using the BERT-base-uncased as the pre-trained MLM for SVM and BiLSTM models [1]. The size of candidate substitutions found by MLM output is set to 100 for all models.Then, W_c is sorted in descending order of importance using Equation 1 and the most important candidate is selected and used to replace w in $S^{(i)}_{rep}$ (line 10).

Algorithm 1 continues in this fashion to generate counterfactual sentences using RM-CT and REP-CT for each sentence in each paragraph of the target document [2]. It returns two counterfactual documents, which correspond to documents produced from the RM-CT and REP-CT sentences; see lines 15–18.

The above approach is not guaranteed to always generate counterfactuals. Typically, reviews that

[1] For Transformers-based models, we use their own pre-trained MLM (e.g., RoBERTa and XLNet) as the generator.

[2] Generating one counterfactual edit for an IMDB instance takes an average of $\approx$ 3.4 seconds based on the RoBERTa-Large model.

cannot be transformed into plausible counterfactuals contain spurious associations that interfere with the model's predictions. For example, in our method, the negative review *"The film is pretty bad, and her performance is overacted"* will be first modified as *"The film is pretty good, and her performance is lifelike"*. The revised review's prediction will remain negative. Meanwhile, the word *"her"* will be identified as a potential causal term. To alleviate this problem, we further conduct the substitution of synonyms for those instances that have been already modified with antonym substitution by using causal terms. As an example, we will continue replacing the word *"her"* with *"their"* until the prediction has been flipped; see also Zmigrod et al. (2019) for related ideas.

In conclusion, then, the final augmented dataset that is produced of three parts: (1) counterfactuals generated by RM-CT; (2) counterfactuals generated by REP-CT; (3) adversarial examples generated by synonym substitutions.

3.3 Ensuring Minimal Changes

When generating plausible counterfactuals, it is desirable to make minimal changes so that the resulting counterfactual is as similar as possible to the original instance (Miller, 2019; Keane and Smyth, 2020). To evaluate this for the approach described we use the MoverScore (Zhao et al., 2019) – an edit-distance scoring metric originally designed for machine translation – which confirms that the MoverScore for the automatic CAD instances is marginally higher when compared to human-generated counterfactuals, indicated greater similarity between counterfactuals and their original instances. The MoverScore between human-generated counterfactuals and original reviews is 0.74 on average (minimum value of 0.55) and our augmented data results in a slightly higher average score than human-generated data for all models. The generated counterfactuals and synonym substitutions that achieve a MoverScore above 0.55 are combined with the original dataset for training robust classifiers.

4 Datasets

Our evaluation uses three different kinds of datasets, in-domain data, challenge data, and out-of-domain data.

State-of-the-art Models	SST-2	IMDB
SMART-RoBERTa (Jiang et al., 2020)	**97.5**	**96.3**
RoBERTa-Large (Liu et al., 2019)	96.7	**96.3**
RTC-attention (Zhang and Zhang, 2019)	90.3	88.7
Bi-LSTM	86.7	86.0

Table 1: The performance of state-of-the-art models in sentiment analysis.

4.1 In-domain Data

We first adopt two of the most popular benchmark datasets – SST-2 and IMDB (Maas et al., 2011) – to show the recent advances on sentiment analysis with the benefit of pre-trained models. However, we mainly focus on the robustness of various models for sentiment analysis in this work, rather than in-domain accuracy. Hence, following Wang and Culotta (2021) and Kaushik et al. (2020), we perform binary sentiment classification experiments on the IMDB dataset sampled from Maas et al. (2011) that contains 1707 training, 245 validation, and 488 testing examples with challenge dataset (paired counterfactuals).

4.2 Challenge Data

Based on the in-domain IMDB data, Kaushik et al. (2020) employ crowd workers not to *label* documents, but to *revise* movie review to reverse its sentiment, without making any gratuitous changes. We directly use human-generated counterfactuals by Kaushik et al. (2020) as our challenge data, enforcing a 50:50 class balance.

4.3 Out-of-domain Data

We also evaluate our method on different out-of-domain datasets, including Amazon reviews (Ni et al., 2019) from six genres: beauty, fashion, appliances, gift cards, magazines, and software, a Yelp review dataset, and the Semeval-2017 Twitter dataset (Rosenthal et al., 2017). These have all been sampled to provide a 50:50 label split. The size of the training data has been kept the same for all methods, and the results reported are the average from five runs to facilitate a direct comparison with baselines (Kaushik et al., 2020, 2021).

5 Results and Discussions

We first describe the performance of the current state-of-the-art methods on sentiment analysis based on the SST-2 and IMDB benchmark datasets. Next, we will discuss the performance benefits by using our automatically generated counterfactuals

Models	**Parameter**	**Training / Testing data**				**AC: (Our method)**			
		O/O	CF/O	CF/CF	O/CF	C/O	AC/O	C/CF	AC/CF
SVM(TF-IDF)	-	80.0	58.3	91.2	51.0	83.7	**84.8**	87.3	86.1
Bi-LSTM	0.2M	79.3	62.5	89.1	55.7	81.5	**82.2**	92.0	88.5
Transformer-based Models									
BERT [ICLR,2021]	110M	87.4	80.4	90.8	82.2	88.5	**90.6**	95.1	92.2
WWM-BERT-Large	335M	91.2	86.9	96.9	93.0	91.0	**91.8**	95.3	94.1
XLNet-Large	340M	**95.3**	90.8	98.0	93.9	93.9	**94.9**	96.9	95.5
RoBERTa-Large	355M	93.4	91.6	96.9	93.0	93.6	**94.1**	96.7	94.3

Table 2: The accuracy of various models for sentiment analysis using different datasets, including the human-generated counterfactual data and counterfactual samples generated by our pipeline. *O* denotes the original IMDB review dataset, *CF* represents the human-revised counterfactual samples, *C* denotes the combined dataset consisting of original and human-revised dataset, and *AC* denotes the original dataset combined with automatically generated counterfactuals. *C* and *AC* contain the same size of training samples (3.4K).

Original Samples	**Original**	**Robust**
Nolan's film...superb directing skills (*POS*)	superb:0.213 film:0.446 Nolan:0.028	0.627 0.019 0.029
It's a poor film, but I must give it to the lead actress in this one (*NEG*)	poor:-0.551 film:-0.257 actress:-0.02	-0.999 -7e-7 -1e-6

Table 3: Less sensitivity to spurious patterns has been shown in the robust BERT-base-uncased model.

on an in-domain test. We further compare our method, human-label method, and two state-of-the-art style-transfer methods (Sudhakar et al., 2019; Madaan et al., 2020) in terms of the model robustness on generalization test. Notably, we provide an ablation study lastly to discuss the influence of *edit-distance* for performance benefits.

5.1 State-of-the-art Models

As the human-generated counterfactuals (Kaushik et al., 2020) are sampled from Maas et al. (2011), the results in Table 1 cannot be directly compared with Table 2 [3]. As shown in Table 1, by comparing BiLSTM to Transformer-base methods, it can be seen that remarkable advances in sentiment analysis have been achieved in recent years. On SST-2, SMART-RoBERTa (Jiang et al., 2020) outperforms Bi-LSTM by 10.8% (97.5% vs. 86.7%) accuracy, where a similar improvement is observed on IMDB (96.3% vs. 86.0%).

According to the results, we select the following models for our experiments, which covers a spectrum of statistical, neural and pre-trained neural methods: SVM (Suykens and Vandewalle, 1999), Bi-LSTM (Graves and Schmidhuber, 2005), BERT-Base (Devlin et al., 2018), RoBERTa-Large (Liu et al., 2019), and XLNet-Large (Yang et al., 2019). The SVM model for sentiment analysis is from scikit-learn and uses TF-IDF (Term Frequency-Inverse Document Frequency) scores, while the Transformer-based models are built based on the Pytorch-Transformer package [4]. We keep the prediction models the same as Kaushik et al. (2020), except for Naive Bayes, which has been abandoned due to its high-variance performance shown in our experiments.

In the following experiments, we only care about whether the robustness of models has been improved when training on the augmented dataset (original data & CAD). Different counterfactual examples have been generated for different models in terms of their own causal terms in practice, while the hyper-parameters for different prediction models are all identified using a grid search conducted over the validation set.

5.2 Comparison with Original Data

On the Influence of Spurious Patterns. As shown in Table 2, we find that the linear model (SVM) trained on the original and challenge (human-generated counterfactuals) data can achieve 80% and 91.2% accuracy testing on the IID hold-out data, respectively. However, the accuracy of the SVM model trained on the original set when testing on the challenge data drops dramatically (**91.2%** vs. **51%**), and vice versa (**80%** vs. **58.3%**). Similar findings were reported by Kaushik et al. (2020), where a similar pattern was observed in the Bi-LSTM model and BERT-base model. This provides further evidence supporting the idea that the spurious association in machine learning models is harmful to the performance on the challenge set for sentiment analysis.

[3] We can only get the human-generated counterfactual examples (Kaushik et al., 2020) sampled from the IMDB dataset.

[4] https://github.com/huggingface/pytorch-transformers

On the Benefits of Robust BERT. As shown in Table 3, we also test whether the sensitivity to spurious patterns has been eliminated in the robust BERT model. We notice that the correlations of the real causal association *"superb"* and *"poor"* are improved from 0.213 to 0.627 and -0.551 to -0.999, respectively. While the correlation of spurious association "film" is decreased from 0.446 to 0.019 and -0.257 to -7e-7 on positive and the negative samples, respectively. This shows that the model trained with our CAD data does provide robustness against spurious patterns.

On the Influence of Model Size. Previous works (Kaushik et al., 2021; Wang and Culotta, 2021) have not investigated the performance benefits on larger pre-trained models. While we further conduct experiments on various Transformer-based models with different parameter sizes to explore whether the larger transformer-based models can still enjoy the performance benefits of CAD (Table 2). We observe that although the test result can increase with the parameter size increasing (best for 94.9% using XLNet), the performance benefits brought by human-generated CAD and the auto-generated CAD declines continuously with the parameter size increase. For example, the BERT-base-uncased model trained on the auto-generated combined dataset can receive 3.2% (90.6% vs. 87.4%) improvement on accuracy while performance increases only 0.6% (91.8% vs. 91.2%) on accuracy for WWM-BERT-Large. It suggests that larger pre-trained Transformer models may be less sensitive to spurious patterns.

5.3 Comparison with Human CAD

Robustness in the In-domain Test. We can see that all of the models trained on automatic CAD – shown as AC in the Table 2 – can outperform the human-generated CAD varying with the models (AC/O vs. C/O) as follows: SVM (+1.1%), Bi-LSTM (+0.7%), BERT-base-uncased (+2.1%), BERT-Large (+0.8%), XLNet-Large (+1.0%), and RoBERTa-Large (+0.5%) when testing on the original data. If we adopt the automatic CAD (AC), we note a distinct improvement in Table 2 across all models trained on the challenge data in terms of *11.3%* in average (AC/O vs. CF/O), whereas the human-generated CAD can achieve 10.2% accuracy improvement (C/O vs. CF/O) in average. It is noteworthy that the human-generated CAD can slightly outperform our method when testing on the human-generated (CF) data, it may be because the training and test sets of the human-generated (CF) data are generated by the same group of labelers.

Out-of-domain Test using Different Training Data	**SVM**	**BERT**
Accuracy on Amazon Reviews		
Orig & CAD (Our Method) (3.4k)	78.6	**84.7**
Orig & CAD (By Human) (3.4k)	79.3	83.3
Orig. & (Sudhakar et al., 2019)	64.0	77.2
Orig. & (Madaan et al., 2020)	74.3	71.3
Orig. (3.4k)	74.5	80.0
Accuracy on Semeval 2017 Task B (Twitter)		
Orig & CAD (Our Method) (3.4k)	69.7	**83.8**
Orig & CAD (By Human) (3.4k)	66.8	82.8
Orig. & (Sudhakar et al., 2019)	59.4	72.5
Orig. & (Madaan et al., 2020)	62.8	79.3
Orig. (3.4k)	63.1	72.6
Accuracy on Yelp Reviews		
Orig & CAD (Our Method) (3.4k)	85.5	**87.9**
Orig & CAD (By Human) (3.4k)	85.6	86.6
Orig. & (Sudhakar et al., 2019)	69.4	84.5
Orig. & (Madaan et al., 2020)	81.3	78.8
Orig. (3.4k)	81.9	84.3

Table 4: Out-of-domain test accuracy of SVM and BERT-base-uncased models trained on the original (Orig.) IMDB review only, Counterfactually Augmented Data (CAD) combining with original data, and sentiment-flipped style-transfer examples.

Robustness in the Generalization Test. We explore how our approach makes prediction models more robust out-of-domain in Table 4. For direct comparison between our method and the human-generated method, we adopt the fine-tuned BERT-base model trained with the augmented dataset (original & automatically revised data). The fine-tuned model is directly tested for out-of-domain data without any adjustment. As shown in Table 4, only our method and the human-label method can outperform the BERT model trained on the original data with average 6.5% and 5.3% accuracy improvements, respectively. Our method also offers performance benefits over three datasets even when compared to the human-label method on BERT.

Neural Method vs. Statistical Method. As shown in Table 4, the performance of the SVM model with automatic CAD is more robust than other automated methods (Sudhakar et al., 2019; Madaan et al., 2020) across all datasets. However, the human-labeled CAD can improve Amazon reviews' accuracy compared to our method using the SVM model by 0.7%. It indicates that human-generated data may lead to more performance benefits on a statistical model.

Types of Algorithms	Examples
Hierarchical RM-CT: Remove negative limitations	Ori: Some films just simply **should not be** remade. This is one of them. In and of itself **it is not a bad film**. Rev: Some films just simply **should be** remade. This is one of them. In and of itself **it is a bad film**.
Hierarchical RE-CT: Replacing the causal terms	Ori: It is **badly directed, badly acted and boring**. Rev: It is **well directed, well acted and entertaining**.
Combined method:	Ori: This movie is **so bad**, it can only be compared to the all-time **worst** "comedy": Police Academy 7. **No laughs** throughout the movie. Rev: This movie is **so good**, it can only be compared to the all-time **best** "comedy": Police Academy 7. **Laughs** throughout the movie.

Table 5: Most prominent categories of edits for flipping the sentiment performed by our algorithms, namely hierarchical RM-CT and hierarchical REP-CT.

5.4 Comparison with Automatic Methods

Automatic CAD vs. Style-transfer Methods. As shown in Table 4, the style-transfer results are consistent with Kaushik et al. (2021). We find that the sentiment-flipped instances generated by style-transfer methods degrade the test accuracy for all models on all kinds of datasets, whereas our method has achieved the best performance for all settings. It suggests that our method have its absolute advantage for data augmentation in sentiment analysis when compared to the state-of-the-art style-transfer models.

Our Methods vs. Implausible CAD. The authors of the only existing approach for automatically generating CAD (Wang and Culotta, 2021) report that their methods are not able to match the performance of human-generated CAD. Our methods consistently outperform human-labeled methods on both *In-domain* and *Out-of-domain* tests. To further provide quantitative evidence of the influence of the *edit-distance* in automatic CAD, we demonstrate an ablation study in Table 6. The result shows that the quality of the generated CAD, which is ignored in the previous work Wang and Culotta (2021), is crucial when training the robust classifiers. In particular, the BERT model fine-tuned with implausible CAD (below the threshold) can receive comparable negative results with the style-transfer samples, alongside the performance decrease on all datasets, except for Twitter.

Training Data	IMDB	Out-of-domain Test		
BERT-base-uncased	Orig.	Amazon	Twitter	Yelp
Orig. & CAD ↑ (3.4K)	**90.6**	**84.7**	**83.8**	**87.9**
Orig. & CAD ↓ (3.4K)	87.1	79.5	73.8	79.0
Orig. (1.7K)	87.4	80.0	72.6	84.3

Table 6: Ablation study on the influence of the *edit-distance* controlled by the threshold of MoverScore. ↑ indicates the CAD (1.7K) above the threshold, while ↓ denotes the CAD (1.7K) below the threshold.

5.5 Case Study and Limitations

The three most popular kinds of edits are shown in Table 5. These are, negation words removal, sentiment words replacement, and the combination of these. It can be observed from these examples that we ensure the edits on original samples should be minimal and fluent as was required previously with human-annotated counterfactuals (Kaushik et al., 2020). As shown in Table 5, we flipped the model's prediction by replacing the causal terms in the phrase *"badly directed, badly acted and boring"* to *"well directed, well acted and entertaining"*, or removing *"No laughs throughout the movie."* to *"Laughs throughout the movie"* for a movie review.

We also noticed that our method may face the challenge when handling more complex reviews. For example, the sentence *"Watch this only if someone has a gun to your head ... maybe."* is an apparent negative review for a human. However, our algorithm is hard to flip the sentiment of such reviews with no explicit casual terms. The technique on sarcasm and irony detection may have benefits for dealing with this challenge.

6 Conclusion

We proposed a new framework to automatically generate counterfactual augmented data (CAD) for enhancing the robustness of sentiment analysis models. By combining the automatically generated CAD with the original training data, we can produce more robust classifiers. We further show that our methods can achieve better performance even when compared to models trained with human-generated counterfactuals. More importantly, our evaluation based on several datasets has demonstrated that models trained on the augmented data (original & automatic CAD) appear to be less affected by spurious patterns and generalize better to out-of-domain data. This suggests there exists a significant opportunity to explore the use of the CAD in a range of tasks (e.g., natural language inference, natural language understanding, and social bias correction.).

Impact Statement

Although the experiments in this paper are conducted only in the sentiment classification task, this study could be a good starting point to investigate the efficacy of automatically generated CAD for building robust systems in many NLP tasks, including Natural Language Inference (NLI), Named Entity Recognition (NER), Question Answering (QA) system, etc.

Acknowledgment

We would like to thank Eoin Kenny and Prof. Mark Keane from Insight Centre for their helpful advice and discussion during this work. Also, we would like to thank the anonymous reviewers for their insightful comments and suggestions to help improve the paper. This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number 12/RC/2289_P2.

References

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Ruth MJ Byrne. 2019. Counterfactuals in explainable artificial intelligence (xai): evidence from human reasoning. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 6276–6282. AAAI Press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Xiao Ding, Dingkui Hao, Yuewei Zhang, Kuo Liao, Zhongyang Li, Bing Qin, and Ting Liu. 2020. HIT-SCIR at SemEval-2020 task 5: Training pre-trained language model with pseudo-labeling data for counterfactuals detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 354–360, Barcelona (online). International Committee for Computational Linguistics.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, et al. 2020. Evaluating models' local decision boundaries via contrast sets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 1307–1323.

Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Counterfactual visual explanations. In *ICML*.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, 18(5-6):602–610.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177.

Alon Jacovi and Yoav Goldberg. 2020. Aligning faithful interpretations with their social attribution. *arXiv preprint arXiv:2006.01067*.

Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. 2020. SMART: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2177–2190, Online. Association for Computational Linguistics.

Xisen Jin, Zhongyu Wei, Junyi Du, Xiangyang Xue, and Xiang Ren. 2019. Towards hierarchical importance attribution: Explaining compositional semantics for neural sequence models. In *International Conference on Learning Representations*.

Jason Jo and Yoshua Bengio. 2017. Measuring the tendency of cnns to learn surface statistical regularities. *arXiv preprint arXiv:1711.11561*.

Rie Johnson and Tong Zhang. 2017. Deep pyramid convolutional neural networks for text categorization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 562–570.

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*.

Divyansh Kaushik, Amrith Setlur, Eduard Hovy, and Zachary C Lipton. 2021. Explaining the efficacy of counterfactually augmented data. In *International Conference on Learning Representations*.

Mark T Keane and Barry Smyth. 2020. Good counterfactuals and where to find them: A case-based technique for generating counterfactuals for explainable ai (xai). In *International Conference on Case-Based Reasoning (ICCBR)*.

Eoin M Kenny and Mark T Keane. 2021. On generating plausible counterfactual and semi-factual explanations for deep learning. In *AAAI*.

Soo-Min Kim and Eduard Hovy. 2004. Determining the sentiment of opinions. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 1367–1373.

Zachary C Lipton. 2018. The mythos of model interpretability. *Queue*, 16(3):31–57.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2020. Gender bias in neural natural language processing. In *Logic, Language, and Security*, pages 189–202. Springer.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, pages 4765–4774.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Aman Madaan, Amrith Setlur, Tanmay Parekh, Barnabas Poczos, Graham Neubig, Yiming Yang, Ruslan Salakhutdinov, Alan W Black, and Shrimai Prabhumoye. 2020. Politeness transfer: A tag and generate approach. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1869–1881, Online. Association for Computational Linguistics.

Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, and Simone Teufel. 2019. It's all in the name: Mitigating gender bias with name-based counterfactual data substitution. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5270–5278.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 188–197.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. 2019. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, pages 13991–14002.

Dino Pedreschi, Fosca Giannotti, Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, and Franco Turini. 2019. Meaningful explanations of black box ai decision systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9780–9784.

Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. 2009. *Dataset shift in machine learning*. The MIT Press.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, pages 502–518.

Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. 2020a. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 180–186.

Dylan Slack, Sophie Hilgard, Sameer Singh, and Himabindu Lakkaraju. 2020b. How much should i trust you? modeling uncertainty of black box explanations. *arXiv preprint arXiv:2008.05030*.

Megha Srivastava, Tatsunori Hashimoto, and Percy Liang. 2020. Robustness to spurious correlations via human annotations. In *International Conference on Machine Learning*, pages 9109–9119. PMLR.

Akhilesh Sudhakar, Bhargav Upadhyay, and Arjun Maheswaran. 2019. "transforming" delete, retrieve, generate approach for controlled text style transfer. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the*

9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3260–3270.

Masashi Sugiyama and Motoaki Kawanabe. 2012. *Machine learning in non-stationary environments: Introduction to covariate shift adaptation*. MIT press.

Johan AK Suykens and Joos Vandewalle. 1999. Least squares support vector machine classifiers. *Neural processing letters*, 9(3):293–300.

Damien Teney, Ehsan Abbasnedjad, and Anton van den Hengel. 2020. Learning what makes a difference from counterfactual examples and gradient supervision. *arXiv preprint arXiv:2004.09034*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.

Ke Wang and Xiaojun Wan. 2019. Automatic generation of sentimental texts via mixture adversarial networks. *Artificial Intelligence*, 275:540–558.

Zhao Wang and Aron Culotta. 2021. Robustness to spurious correlations in text classification via automatically generated counterfactuals. In *AAAI*.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. 2020. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33.

Linyi Yang, Eoin Kenny, Tin Lok James Ng, Yi Yang, Barry Smyth, and Ruihai Dong. 2020a. Generating plausible counterfactual explanations for deep transformers in financial text classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6150–6160.

Xiaoyu Yang, Stephen Obadinma, Huasha Zhao, Qiong Zhang, Stan Matwin, and Xiaodan Zhu. 2020b. SemEval-2020 task 5: Counterfactual recognition. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 322–335, Barcelona (online). International Committee for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5753–5763.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489.

Yuan Zang, Fanchao Qi, Chenghao Yang, Zhiyuan Liu, Meng Zhang, Qun Liu, and Maosong Sun. 2020. Word-level textual adversarial attacking as combinatorial optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6066–6080.

Huangzhao Zhang, Hao Zhou, Ning Miao, and Lei Li. 2019. Generating fluent adversarial examples for natural languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5569.

Yuan Zhang and Yue Zhang. 2019. Tree communication models for sentiment analysis. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3518–3527.

Yue Zhang, Qi Liu, and Linfeng Song. 2018. Sentence-state lstm for text representation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 317–327.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578.

Ran Zmigrod, Sebastian 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.