Making a (Counterfactual) Difference One Rationale at a Time

Anonymous Author(s)
Affiliation
Address
email

Abstract

Rationales, *snippets of extracted text* that explain an inference, have emerged as a popular framework for interpretable natural language processing (NLP). Rationale models typically consist of two cooperating modules: *a selector* and *a classifier* with the goal of maximizing the mutual information (MMI) between the "selected" text and the document label. Despite their promises, MMI-based methods often pick up on spurious text patterns and result in models with nonsensical behaviors. In this work, we investigate whether *counterfactual data augmentation (CDA)*, without human assistance, can improve the performance of the selector by lowering the mutual information between spurious signals and the document label. Our counterfactuals are produced in an unsupervised fashion using class-dependent generative models. From an information theoretic lens, we derive properties of the unaugmented dataset for which our CDA approach would succeed. The effectiveness of CDA is empirically evaluated by comparing against several baselines including an improved MMI-based rationale schema [16] on two multi-aspect datasets. Our results show that CDA produces rationales that better align with the signal of interest.

1 Introduction

Research in neural model interpretability has been cast as important and received significant recent attention [18]. Within the field of natural language processing (NLP), *rationales* have been a popular method for providing interpretability in the form of extracted subsets of text [10]. Rationale models typically consist of two cooperating modules where one module, the 'rationale selector', selects the rationale from a source document, and the other module, the 'classifier', acts on only the selected rationale without seeing the rest of the document. There is interpretability through sparsity and exclusivity.

A common approach for training these rationale models is based on the maximum mutual information criteria (MMI) [7]. With the MMI criteria, rationale selectors seek the subset of text that carries the most information about the target label. Often, sparsity and coherency constraints are used to keep the rationales interpretable. Within many datasets, however, *spurious patterns* and *co-varying aspects* can cause the rationale selector to pick up on patterns that do not capture a desired relationship between input text and target labels. As a result, the rationalized model can have undesirable behaviour like predicting a hotel is very clean because it is in a convenient location. Nonsensical rationales or explanations might decrease trust in the model, and in some cases, suggest the model might generalize poorly [22].

In this work, we propose a general *counterfactual data augmentation (CDA) [19] approach* to aid rationale models trained with MMI. We show that theoretically our CDA approach can effectively improve the performance of rationale selectors by lowering the mutual information between spurious

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

signals and aspects of interest. Empirically, we show that models trained on our CDA datasets learn higher quality rationales than those trained on the original dataset when both use the same MMI criteria. More importantly, the most significant advantage of our CDA approach is that it does not require human intervention. We use rationales from an initial, noisy model and replace them with new text that changes the target label using a generative neural model. In this way, our CDA approach is completely hands off and does not need input from human experts or crowd workers.

We first show that in the extremely ideal scenario where the initial rationale selector is perfect, our CDA approach can eliminate the mutual information between spurious signals and the target label. Next, we show in the common, realistic scenario where the rationale selector is noisy and imperfect, our CDA approach can still yield gains. Finally, the effectiveness of our CDA approach is compared against several baselines including an improved MMI-based rationale schema [16] on multi-aspect review datasets, *TripAdvisor* [25] and *RateBeer* [21], which are commonly used in this research area.

2 Counterfactual Data Augmentation and Multi-Aspect Datasets

2.1 Definitions and Notations

We use upper-case letters to denote random variables, X and Y, and lower-case letters to denote samples from these variables, x and y. $I(X, Y)$ marks the mutual information between X and Y. Mutual information is defined as the reduction in uncertainty of a random variable due to knowledge in another random variable, $I(X, Y) = H(Y) - H(Y|X)$, where $H(X)$ and $H(Y|X)$ are the Shannon entropy and conditional entropy respectively [8].

2.2 Problem Formulation

This work follows the same rationale concept introduced in [16]. Specifically, one neural module extracts text from a document and another neural module classifies the extracted text. Later it has been shown that the rationalization criteria "aims to maximize the mutual information between the response variable from the model and selected features" [7]. That is,

$$\max_{G} I(X_M; Y) \quad \text{subject to} \quad M \sim G(X) \tag{1}$$

where M denotes a binary rationale mask over the input produced by a rationale selector G. This states that the goal of the rationale selector is to select the subset of features in X that are *most informative* of the the label Y under some constraints defined by the selector G. In this work, our selector restraint is the size of the rationale as a fixed percentage of the input text.

Here we consider a multi-aspect dataset, D, with features *X* and labels *Y*. X is a set of features, or a sequence of words in NLP, and Y is a vector, possibly one dimensional, of numerical scores. In a multi-aspect dataset, a single document can discuss multiple attributes of a single object. For example, a single beverage review might discuss its appearance, taste, and smell. We assume some subset of the features, X_1, belong to the target aspect label, Y_1, while other features, X_2, are spurious. These features could belong to other aspects or be artifact of the dataset. In the following, for simplicity, we will use <X, Y_1> or <X_1, X_2, Y_1> interchangeably based on the context.

Our goal is to estimate the score for aspect Y_1 while only using X_1 as the rationale. Our model to predict Y_1 should be some function, f that depends only on X_1. $Y_1 = f(X_1)$. For each $< X_1, Y_1 >$ pair, there is one corresponding model. In this work, we use multi-aspect datasets, and model each aspect individually. This simulates the more general case where a dataset has a single output label of interest and all signals in the dataset that do not pertain to that label are considered spurious or belonging to other, not-estimated aspects.

When predicting Y_1, our ideal model would focus on X_1 and ignore all other features in X_2. Ideally, for a given sample <x, y>, our selector would select a subset of x, $x_1 = x_M$. There are two primary reasons a rationale selector might converge to point where it fails to effectively extract x_1. First, [7] showed finding a direct solution to Eqn 1 is intractable and they derive a variational approximation with Monte Carlo based gradient estimates. Second, datasets can contain artifacts such that spurious patterns might contain significant mutual information with the target label. Along the same lines but even more concerning, other aspects in a dataset might be highly correlated with the target label. For example, beers that smell good usually taste good as well.

Because the selector will naturally select features that maximize $I(X_M, Y_1)$, we can help the selector find X_1 by lowering the mutual information between the other spurious features and the desired label, $I(X_2, Y_1)$. Drawing on ideas from [19], we do this through a general counterfactual data augmentation (CDA) scheme where, in the counterfactual dataset, superscript c, we flip the label of the document from Y_1 to Y_1^c and replace the text selected by the rationale selector, X_1, with an inference, X_1^c, generated by a *class conditioned masked language model (MLM)* using Y_1^c and X_2 described as:

$$Y_1^c \leftarrow 1 - Y_1; \quad X_1^c \leftarrow \arg\max_{X_1} p(X_1|1 - Y_1, X_2) \tag{2}$$

In our generated counterfactual dataset $< X_1^c, X_2, Y_1^c >$, X_1^c is the newly generated counterfactual, X_2 is the original spurious feature set, and we assign $Y^c = 1 - Y_1$. We will show, in the augmented dataset which is the concatenation of datasets $< X_1, X_2, Y_1 >$ and $< X_1^c, X_2, 1 - Y_1 >$, we are lowering the mutual information $I(X_2, Y_1)$. We will use superscript a for the augmented dataset: D^a = <X_1^a,X_2^a, Y_1^a>.

2.3 Lowering $I(X_2, Y_1)$, an idyllic case

Consider a dataset of beer reviews where each document contains a description of the taste and smell of a beer as well as a numerical score for both aspects. The task is to estimate the smell score while using the smell text as the rationale. Figure 1 demonstrates our process. An example of a concise document in this dataset might be *This beer smells great. It tastes terrific*. In this document x_1 is the phrase *smells great* and x_2 is *tastes terrific*. Both have positive sentiment but our only label for this document is $y_1 = 1$ for the smell sentiment. Our counterfactual document could be *This beer smells awful. It tastes terrific* and our label becomes $y_1 = 0$. Through this simple example of comparing the original document with its counterfactual, we can see that in the augmented dataset $p(Y_1|X_2) = p(Y_1)$ so $I(X_2, Y_1)$ is 0 and $I(X_1, Y_1)$ is unchanged.

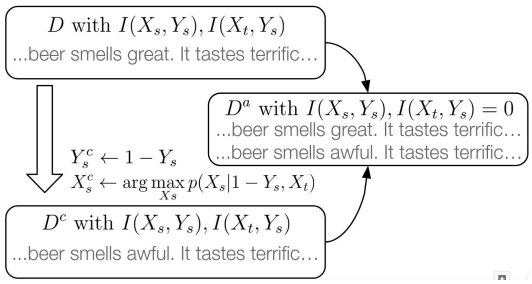

Figure 1: Toy example to demonstrate our approach. In the augmented dataset D^a, the mutual information, I, between the smell score and the smell text is preserved while the mutual information between the smell score and the taste text is eliminated.

With perfect knowledge of the ground truth rationales and the process $X_1^c \leftarrow p(X_1|1 - Y_1, X_2)$, we can craft counterfactual documents and therefore a counterfactual dataset that perfectly eliminates $I(X_2, Y_1)$ while preserving $I(X_1, Y_1)$. However, the challenge is that ground truth rationales are not provided in the training data. Therefore, it is important for us to show that even when rationales selected by the initial rationale selector are noisy and imperfect, we can still lower $I(X_2, Y_1)$ in the augmented dataset and benefit subsequent models trained with MMI.

2.4 Dealing with a Noisy Initial Selector

Here we are working on the augmented dataset: D^a = <X_1^a, X_2^a, Y^a>. To completely eliminate $I(X_2^a, Y_1^a)$, our CDA approach requires a perfect rationale selector. We were originally motivated by improving a poor rationale selector, so here we track what happens to both $I(X_2^a, Y_1^a)$ and $I(X_1^a, Y_1^a)$ when the rationale selector is not perfect. Our new goal is to reduce $I(X_2^a, Y_1^a)$ more than we reduce $I(X_1^a, Y_1^a)$. We analyze the procedure in the worst case scenario in order to create a lower bound on our algorithm's benefits under some assumptions.

In an extremely erroneous case, we say that for a given <x_1, x_2, y_1>, our initial rationale selector mistakenly selects x_2 when aiming for x_1. When creating the corresponding counterfactual document, we still have $y_1^c = 1 - y_1$, but we modify the document according to $x_2^c \leftarrow \arg\max_{X_2} p(x_2|1 - y_1, x_1)$ instead of the process defined by Eqn 2. Going back to our concise example, this would be the counterfactual document *This beer tastes good. It also smells bad*. In this extremely erroneous case, we have decreased $I(X_1, Y_1)$ while $I(X_2, Y_1)$ remains unchanged. Thus, we define the worst case scenario as reducing $I(X_1, Y_1)$ at some error rate α and keeping $I(X_2, Y_1)$ constant at the same rate.

If we say that this error happens to all samples with rate α, we can analyze conditions that must be present in the original dataset so that our CDA approach is beneficial. Lets first define ΔI^a as the change in mutual information from the original to the augmented dataset.

$$\Delta I^a_{X_i,Y_j} = I(X_i, Y_j) - I(X^a_i, Y^a_j) \tag{3}$$

In order for our CDA approach to be beneficial, we need to decrease $I(X_2, Y_1)$ more than $I(X_1, Y_1)$. This is the relation

$$0 < \Delta I^a_{X_2,Y_1} - \Delta I^a_{X_1,Y_1} \tag{4}$$

We make the approximation that $p(X_1) = p(X_1^{er}) = p(X_1^c)$ where $p(X_1^{er})$ is the X_1 that occurs when we make an error. We also assume $p(Y_1^c|X_1^c) = p(Y_1|X_1)$. These are especially safe assumptions when $p(Y_1) = \frac{1}{2}$ which is the common case of a balanced dataset. Note that $p(Y_1|X_1^{er}) = p(Y_1)$. We now have a mixture of distributions in the augmented dataset.

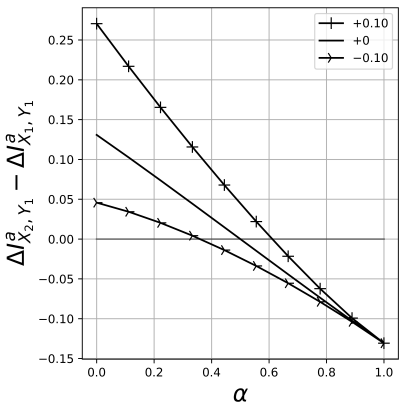

Figure 2: CDA benefits (y-axis) as function of error rate α. Each line represents $p(Y_1|X_2) = p(Y_1|X_1) \pm c$.

$$p(Y_1^a|X_1^a) = \alpha p(Y_1) + (1-\alpha)p(Y_1|X_1) \tag{5}$$

For X_2, we have the reverse.

$$p(Y_1^a|X_2^a) = (1-\alpha)p(Y_1) + \alpha p(Y_1|X_2) \tag{6}$$

We can now expand Eqn 4 using the definition $I(X,Y) = H(Y) - H(Y|X)$.

$$0 < -H(Y_1|X_2) + H(Y_1^a|X_2^a) + H(Y_1|X_1) - H(Y_1^a|X_1^a) \tag{7}$$

Using the definition, $H(Y|X) = -E\ \log p(Y|X)$, we can expand this further to

$$0 < -E\ \log p(Y_1|X_1) + E\ \log p(Y_1^a|X_1^a) + E\ \log p(Y_1|X_2) - E\ \log p(Y_1^a|X_2^a) \tag{8}$$

Eqn 8 expresses conditions that must be met in our original dataset in order for our CDA procedure to yield gains for some error rate α. We cannot use this relation directly because it requires exact knowledge of the ground-truth rationales.

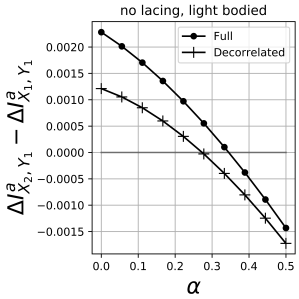

Figure 3: CDA benefits when approximating X_1 and X_2 as the occurrence of bi-grams in documents.

Figure 2 shows the efficacy of the CDA approach when we approximate X_1 and X_2 with binary variables, $p(Y_1|X_1) = .75$, $p(X_1) = p(X_2) = \frac{1}{2}$, and $p(Y_1) = .5$. When X_1 and X_2 are equally informative of Y_1, the benefits of CDA decrease linearly with the error rate, and intuitively, our error rate must be less than 50% to see gains. When X_2 is more informative than X_1, we have a higher error budget to see any benefit, and when X_1 is more informative than X_2, our error budget is smaller. This analysis shows that generally, when the spurious signals have a similar amount of information, we have a higher error budget. If the spurious signals carry much less information, the initial selector needs a small error rate to see any benefit from the CDA approach.

For the datasets used in this work, we can gain insight by approximating X_1 and X_2 with guessable bigrams. For example, we can make another concise example and assume that X_1 is a binary variable that indicates the occurrence of the phrase *no lacing* and X_2 is another binary variable for the phrase *light bodied*. When estimating the appearance aspect for a beer, *no lacing* is a strong indicator for a negative score, and when estimating the palate aspect *light bodied* is also a strong negative indicator. Because the appearance of beer is highly correlated with palatability, both phrases are indicators of a poor appearance score. Taking Y_1 as the appearance score, X_1 as the occurrence of *no lacing*, and *light bodied* as X_2, we can numerically use Eqn 8 to see how CDA helps us for varying error rates in Figure 3. CDA is helpful whenever the curve is above zero. For the fully correlated dataset, described in section 4.1, the information carried by *light bodied* is closer to that of *no lacing* than it is in the decorrelated dataset. We therefore have a higher error budget in the full dataset and more opportunity for the our CDA approach.

3 Architecture and Implementation

3.1 Rationale Framework

The original rationale framework, as viewed in this work, was introduced by [16]. It was not our goal to change the core rationalization algorithm, so we re-implemented the algorithm and updated some details. At a high level, our rationale framework is the same in that we use one network to select the rationale and another network to classify the text. The rationale framework can be visualized by the blue portion of Figure 4.

The original implementation [16] used RNNs for both networks, REINFORCE [26] for dealing with the discontinuity introduced by the binary rationale mask, and variable percentage rationales. In this work, we use transformers for both networks because of their effectiveness over RNNs in NLP [24], the simpler straight-through method [2] [5] instead of REINFORCE, and we use fixed percentage rationales because it eliminates sparsity related hyper-parameters. Our fixed percentage rationales differ from [7] in that we use the top-K tokens during training and inference whereas [7] uses an iterative re-sampling approach with Gumbel-softmax reparameterization [13].

The classifier is trained only to make quality predictions against the labels while using the rationale. This is the cross-entropy between the labels and classifier's prediction, L_y. We follow [16] by using a coherency regularizer for the rationale selector: $L_r = \frac{\lambda_r}{T} \sum_{1 \ldots T} |m_t - m_{t-1}|$ where T is the total number of tokens in a document, m is the rationale mask, and λ_r is the hyper-parameter used to tune coherency. This encourages the rationales to be contiguous. For the datasets evaluated in this work, coherency is a useful inductive bias. The loss for the rationale selector is the coherency regularizer and the cross-entropy between the labels and the classifier's prediction. This is $L_s = L_r + L_y$.

3.2 Counterfactual Predictor

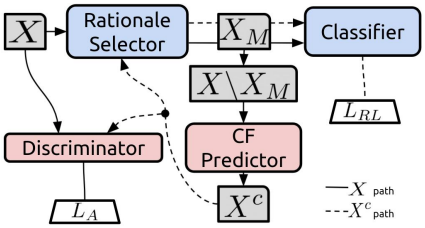

Figure 4: CF Predictor training flow. This procedure is repeated for a class-0 and a class-1 model separately.

The CDA process presented earlier, Eqn 2, requires us to generate new documents with the label and text flipped for only one aspect. We are sampling a new document from $p(X_1|X_2, 1 - Y_1)$. The core challenge is that we are not provided with ground truth rationales or counterfactuals from which to learn this data generating process. This can be viewed under the lens of unsupervised style transfer for which there is significant prior work in the NLP domain [17] [20]. Our method for generating counterfactual documents leverages many ideas from these works, and our main contribution here is connecting these ideas to the rationale framework. We incorporate the rationale framework in the counterfactual generation process because our goal is to lower the mutual information between the spurious signals and the target label. An off-the-shelf style transfer method might focus on signals other than that selected by our initial rationale selector.

The CF Predictor, Figure 4, replaces rationale tokens with tokens predicted by class-dependent masked-language-models (MLMs) [9]. These class dependent MLMs are trained to produce documents with the desired class label through reinforcement learning in a manner similar to [12], and they are trained to generate realistic documents through adversarial training [11]. We use straight-through [2] to propagate gradients through token selection during training. For our datasets with binary labels, we have a MLM model for generating class-0 documents and another MLM model for generating class-1 documents for a single target aspect [27]. The loss for the counterfactual predictor is

$$L_{CFP} = \lambda_{RL} L_{RL} - \lambda_A L_A \tag{9}$$

L_{RL} is provided by our pre-trained and frozen rationale framework. Specifically, this loss is the cross-entropy between the desired label, all ones or all zeros, and the predicted label after passing the counterfactual through the rationale selector and the classifier, $F(G(x^c))$. Where x^c is the generated counterfactual document.

For adversarial training [11] [28], we have a discriminator, $\mathcal{D}$, seeking to distinguish generated documents from the originals. The discriminator's loss, L_D, is $\frac{1}{\lambda_A} L_A$ where L_A is the cross-entropy

Label	Document
negative	**appearance** : *golden clear white* **quickly dissipating head leaves no** *lacing left on the glass . aroma : cereal grains corn and rice adjunct ridden mass produced beer . taste : off adjunct flavor light sweetness no hops evident not real good . mouthfeel : light bodied over carbonated fizzy yellow beer . drinkability : not the beer i would drink .*
positive	**appearance** : *golden clear white* **1 finger head . excellent** *lacing left on the glass . aroma : cereal grains corn and rice adjunct ridden mass produced beer . taste : off adjunct flavor light sweetness no hops evident not real good . mouthfeel : light bodied over carbonated fizzy yellow beer . drinkability : not the beer i would drink .*

Figure 5: Subset of the augmented dataset for beer appearance. The rationale is bold in the original document (top). The replaced words are bold in the counterfactual document (bottom). x_1 and y_1 have changed from original to counterfactual. Ground truth rationales and counterfactuals are not provided in any training data.

between real-fake labels and the prediction given by the discriminator. We include $\frac{1}{\lambda_A}$ when training the discriminator to hamstring it relative to the predictor. We found that this generally helped us avoid mode collapse commonly seen when adversarially training generative models. For L_A, we mask the contribution of original documents without the desired label. When training the class-1 counterfactual predictor, we show the discriminator real documents X with $Y_1 = 1$ and counterfactual documents, X^c, where the original documents' labels were $Y_1 = 0$.

During training, the counterfactual is produced in one step. All words not included in the original rationale, X_M, are kept in the counterfactual document. The kept tokens are $X \backslash X_M$. All words in the original rationale are replaced by the CF Predictor using one prediction from the MLMs in a greedy fashion. After training, when producing the counterfactual dataset, the counterfactual documents still keep all non-rationale tokens. We now replace the rationale tokens from left to right using the output of the CF Predictor MLMs. A counterfactual token at position t is decoded according to the following process.

$$x^c_{1,t} = \arg\max_{x_{1,t}} p(x_{1,t} | x_2, 1 - y_1, x^c_{1,0\ldots t}) \tag{10}$$

We found this to be a good trade-off between greedy decoding and a more expensive beam search. Greedy decoding could generate frequent, repeated tokens. Beam search was an unnecessary expense for generating documents that reflect the target distribution, but do not necessarily need to pass the bar of human readers.

Figure 5 shows an example original document and counterfactual. Notice this example goes back to our *no lacing* and *light bodied* check from Figure 3. *no lacing* becomes *excellent lacing* while *light bodied* is unchanged. With this pair of original and counterfactual documents, we have decreased the mutual information between the phrase *light bodied* and the appearance score.

4 Experiments

4.1 Datasets

We conduct experiments using two datasets. The first is a collection of reviews collected by [25] from *TripAdvisor.com*. We use the training, dev, and test set curated by [1] and used for rationalization by [5] but modify them so that the dev and test samples from all aspects do not appear in any of the training sets. We focus on the location aspect.

The second consists of reviews collected by [21] from *RateBeer*. Each review is a paragraph of text with five numerical scores in the range of 1 to 5 for the appearance, smell, palate, and overall aspects of the beer. We use procedurally decorrelated subsets of the data created by [16] for the appearance, smell, and palate aspects. Following [5], we binarize this data so that all reviews with a score ≥ 3 are class 1 and all reviews with a score ≤ 2 are class 0. We then balance the dataset between classes.

We also present results for the *RateBeer* dataset where the aspects are not decorrelated. Here, we use all available data after binarizing using the same scheme, trimming the dataset so that they

are the same size as the decorrelated sets, and balancing the data between classes. Appendix 1 shows correlation matrices before and after decorrelating the data. In this version of the dataset, the correlations and presumably the mutual information between aspects is much higher than in the decorrelated datasets. This makes the rationalization task more difficult. We believe this is the first work to present rationale results on data that follows the distribution of the full *RateBeer* dataset.

4.2 Baselines

We use three baselines. The first, 'MMI', is the original Rationalization scheme from [16] but re-implemented and updated as described in section 3.1. This model is trained on the unaugmented dataset.

The second, 'Aug.', is implemented to establish our gains are not due only to data augmentation but due to the presented CDA procedure. We augment the original dataset with new samples as generated by the CF predictor models but instead of flipping the label and passing it to the $1 - y_1$ CF Predictor component, we pass the sample to the component of the CF predictor with the same label as the original document. This is factual data augmentation. These new samples are generated by the process

$$y_1^c \leftarrow y_1; \qquad x_1^c \leftarrow \arg\max_{x_1} p(X_1|y_1, x_2) \tag{11}$$

Our third baseline, 'A.CDA.', is a counterfactual data augmentation scheme that does not use neural models. When generating the counterfactual, we replace words in the rationale with antonyms as provided by [4]. We only accept antonyms that are in the vocabulary used by the models and antonyms that have the same part of speech as labeled by NLTK [3].

4.3 Experiment Settings and Asssumptions

We set the rationale percentage to 10% for all datasets. We train the rationale selector and the classifier together, early stop based on the selector cost, freeze the selector, and finally fine tune the classifier on the original dataset. For all of the data sets and models, we use the dev set for early stopping and please see Appendix 2 for details. We pretrain our MLM transformers [24] on unlabeled data from the *RateBeer* and *TripAdvisor* datasets separately. For the TripAdvisor dataset, we pretrain on all data that does not appear in any dev or annotated dataset from [5]. For the *RateBeer* datasets, we pretrain on all data that does not appear in any train, dev, or annotated dataset from [16]. For the *TripAdvisor* dataset, we follow the same strategy subtracting dev and annotated data from [5]. We used a masking rate of 10% and masked tokens were treated as described in [9]. The rationale selector, classifier, CF predictor components, and the GAN discriminator are all initialized from the same MLMs (within a dataset). These models all have a vocabulary with 2^{15} tokens, 8 layers, 8 attention heads, and a hidden dimension of 256. Appendix 3 shows our server configurations and more details on our experiment setup.

Models are selected and reported based on the best performance on the dev set across a grid search. All methods are evaluated using the same grid search. The initial selector used to train the counterfactual predictors was the selected MMI model. The parameters and checkpoints for the CF Predictor models are tuned and chosen to maximize the accuracy of the training documents' predicted label as compared to the target label (measured by the original rationale model) and to maximize the entropy in the inserted counterfactual tokens.

4.4 Results

We evaluate the rationale models by the precision (P), recall (R), and F1 scores of the rationale selectors as compared to human annotations. These token-level metrics are taken as the mean across samples in the annotated set. We also report the accuracy of the classifier on the development set (D.A.). As shown in Tables 1, 2, and 3, our CDA approach outperforms all baselines on $\frac{6}{7}$ experiments.

The improvements in F1 score were least notable for the *appearance* aspect in the decorrelated Beer dataset in Table 2. For this dataset, the initial MMI selector had a low error rate probably because the dataset is already decorrelated. Thus our CDA approach has less potential for improvement. The one experiment for which our CDA approach failed was the palate aspect in the correlated

RateBeer dataset. Notice here that the initial MMI selector was the worst across all experiments for the precision metric. Going back to our analysis in Figure 2, it is likely the initial selector was sufficiently noisy such that $I(X_1, Y_1)$ was lowered more than $I(X_2, Y_1)$ was lowered. For this dataset, our counterfactual predictors had a difficult time flipping the label of the counterfactual document as evaluated by the initial selector-classifier pair. Because of this, the factual augmentation strategy was the most successful for this dataset. For all other datasets, the improvements in the quality of the rationale selector is noticeable.

The baseline augmentation schemes (Aug., A.CDA) had mixed results. The factual augmentation scheme should not have changed the mutual information between the other aspects and the target label. For cases where it performed much worse than MMI alone, it most likely only introduced noise into the dataset. The antonym counterfactual data augmentation strategy, might be effective at lowering the mutual information between the other aspects and the target label but is not sampling from $p(x_1|1 - y_1, x_2)$, so we expect it to also lower $I(X_1, Y_1)$.

Table 1: TripAdvisor - Location

	D.A.	P	R	F1
MMI	84.3	44.2	56.4	43.7
Aug.	84.4	43.2	54.4	42.4
A.CDA.	72.3	9.5	9.2	8.3
CF.CDA	84.3	**47.9**	**60.4**	**47.3**

Empirically, we have shown that models trained on the CDA augmented data tend to outperform models trained on the original datasets. This lends credence to the idea that our scheme is indeed lowering the mutual information between the other aspects and the target label.

Case Study: To illustrate their performance, Figure 6 presents a case study comparing the rationales selected by MMI and our proposed CDA approach in the TripAdvisor and Beer dataset respectively. As shown in Figure 6, our CDA approach can select text that better aligns with human annotations. These models better avoid selecting sentiment-carrying text not pertaining to the aspect of interest.

Table 2: Decorrelated RateBeer Results

	Appearance				Smell				Palate			
	D.A.	P	R	F1	D.A.	P	R	F1	D.A.	P	R	F1
MMI	82.0	91.9	54.7	65.9	84.5	66.4	47.2	52.8	75.5	53.9	50.8	49.2
Aug.	73.1	55.0	31.3	38.4	83.2	75.4	54.0	60.3	75.5	57.9	54.4	52.9
A.CDA.	81.1	89.8	53.1	64.1	72.5	10.2	7.0	7.9	61.4	53.0	48.9	48.0
CF.CDA	81.6	**92.9**	**55.5**	**66.8**	83.5	**76.5**	**54.8**	**61.1**	72.3	**62.0**	**57.2**	**56.1**

Table 3: Correlated RateBeer Results

	Appearance				Smell				Palate			
	D.A.	P	R	F1	D.A.	P	R	F1	D.A.	P	R	F1
MMI	77.1	49.9	28.8	35.1	79.1	52.4	37.2	41.6	83.8	39.7	36.9	35.9
Aug.	64.8	53.4	31.1	37.7	84.3	52.7	36.8	41.5	84.6	**46.9**	**43.7**	**42.6**
A.CDA.	67.5	56.9	33.1	40.2	66.6	4.1	3.0	3.3	74.5	24.7	21.9	21.9
CF.CDA	66.0	**58.1**	**33.9**	**41.1**	79.9	**58.9**	**41.9**	**46.8**	83.9	31.7	28.8	28.4

5 Related Work

The rationale framework used in this work was introduced in [16] and connected to the MMI criteria by [7]. Much of the follow-up work introduced modifications to the learning algorithm to overcome spurious signals picked up by MMI. [29] sought to limit the signal left in the complement of the rationales. The concept of class-wise rationalization was introduced in [5], and [6] introduced a rationale algorithm that seeks invariant rationales across environments. These works modify the rationale framework and the MMI criteria to help the rationale selector find desirable signals. In our work, we keep the MMI criteria unchanged, but instead seek to diminish the undesirable signals through counterfactual data augmentation.

Data augmentation generally and counterfactual data augmentation (CDA) specifically has been a popular technique in recent natural language processing (NLP) work. [19] introduced counterfactual

Hotel - Location
a nice and clean hotel to stay for business and leisure . but the location is not good if you need public transport . it took too long for transport and waiting for bus . but the swimming pool looks good although i never tried .

Decorrelated - Palate
poured from the caged adn corked bottle an almost dull yellow , cloudy but with a large foamy head sweet citrus and malt , floral and herbal notes in the nose . yeast and bread also . very pleasant . the taste was dominated by the sweet and citrus , with a good balance of bready/yeasty taste . not very subtle , the taste of this saision can definitely be described as 'big '. medium to light bodied , the sweet seemed to outweigh any other feel and faded too quickly drinkable for sure but not a fantastic farmhouse; a good sharing beer and a great effort on victory 's part .

Figure 6: Examples from the annotated sets. Hotel-Location (top) and Decorrelated-Palate (bottom). *Human annotations are underlined*. CDA rationales are in blue. MMI rationales are in red. Overlaps between CDA and MMI are in magenta.

data augmentation as a general methodology and showed their rule based scheme could mitigate gender biases commonly seen in word embeddings and NLP models. Other works used CDA to train better named entity recognition (NER) models in the medical text domain [30].

More work [14] continued the ideas of CDA by working with sentiment analysis and natural language inference tasks. They constructed their CDA datasets using crowd workers and showed models trained on the augmented datasets generalized better out-of-domain. Follow up work [15] showed how their crowd worker interventions reduce spurious signals when viewed in a simplified causal modeling lens. They also showed unsupervised text style transfer methods could not produce a counterfactually augmented dataset where a model trained on that data performs as well as training on the original data only in the out-of-domain setting. The causal lens and generative counterfactual data augmentation has also shown recent, positive results in the computer vision domain [23].

6 Conclusions

This work presents a counterfactual data augmentation method for lowering the mutual information between spurious signals and a target label in a dataset. We derive theoretical conditions on the dataset that indicate when our approach will be beneficial. Empirically, we show rationale models trained with MMI on our counterfactually augmented datasets result in rationale selectors with improved behaviours. Our method does not make any changes to the core rationale algorithm suggesting our method could work with other strategies as long as they produce rationales. Furthermore, our analysis shows that the benefits of our scheme are proportional to the error rate of the original rationale selector. This suggests we could perform counterfactual data augmentation iteratively where we achieve better rationale selectors after each iteration that could be used as the initial selector in the next round of data augmentation.

7 Broader Impact

This work focuses on methods for training interpretable NLP models. While interpretability is typically a desired trait, defining exactly what is interpretable is not always clear [18]. Our method uses data augmentation to lower the mutual information between some signals in documents and a label of interest. When the initial rationale selector aligns with human judgment, models trained on the augmented data tend to better align with human judgement. The method can be thought of as amplifying signals relative to others. We are changing the properties of the dataset. For sensitive tasks, it might be necessary to check for undesired biases and fairness concerns both before and after CDA. Along the same lines, it might be possible that this scheme could be used with a human in the loop to lower signals in a dataset which are undesirable. This could be used for making a dataset more fair. A similar rule-based CDA approach has already been used for this purpose [19].

References

[1] Yujia Bao, Shiyu Chang, Mo Yu, and Regina Barzilay. Deriving Machine Attention from Human Rationales. *arXiv:1808.09367 [cs]*, August 2018. arXiv: 1808.09367.

[2] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or Propagating Gradients Through Stochastic Neurons for Conditional Computation. *arXiv:1308.3432 [cs]*, August 2013. arXiv: 1308.3432.

[3] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

[4] Pradipta Bora. geekpradd/PyDictionary, May 2021. original-date: 2014-12-26T10:21:02Z.

[5] Shiyu Chang, Yang Zhang, Mo Yu, and Tommi S. Jaakkola. A Game Theoretic Approach to Class-wise Selective Rationalization. *arXiv:1910.12853 [cs, stat]*, October 2019. arXiv: 1910.12853.

[6] Shiyu Chang, Yang Zhang, Mo Yu, and Tommi S. Jaakkola. Invariant Rationalization. *arXiv:2003.09772 [cs, stat]*, March 2020. arXiv: 2003.09772.

[7] Jianbo Chen, Le Song, Martin J. Wainwright, and Michael I. Jordan. Learning to Explain: An Information-Theoretic Perspective on Model Interpretation. *arXiv:1802.07814 [cs, stat]*, June 2018. arXiv: 1802.07814.

[8] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA, 2006.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[10] Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A Benchmark to Evaluate Rationalized NLP Models. *arXiv:1911.03429 [cs]*, April 2020. arXiv: 1911.03429.

[11] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks. *arXiv:1406.2661 [cs, stat]*, June 2014. arXiv: 1406.2661.

[12] Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. Toward Controlled Generation of Text. *arXiv:1703.00955 [cs, stat]*, September 2018. arXiv: 1703.00955.

[13] Eric Jang, Shixiang Gu, and Ben Poole. Categorical Reparameterization with Gumbel-Softmax. *arXiv:1611.01144 [cs, stat]*, August 2017. arXiv: 1611.01144.

[14] Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. Learning the Difference that Makes a Difference with Counterfactually-Augmented Data. *arXiv:1909.12434 [cs, stat]*, February 2020. arXiv: 1909.12434.

[15] Divyansh Kaushik, Amrith Setlur, Eduard Hovy, and Zachary C. Lipton. Explaining The Efficacy of Counterfactually-Augmented Data. *arXiv:2010.02114 [cs, stat]*, October 2020. arXiv: 2010.02114.

[16] Tao Lei, Regina Barzilay, and Tommi Jaakkola. Rationalizing Neural Predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Austin, Texas, 2016. Association for Computational Linguistics.

[17] Juncen Li, Robin Jia, He He, and Percy Liang. Delete, Retrieve, Generate: A Simple Approach to Sentiment and Style Transfer. *arXiv:1804.06437 [cs]*, April 2018. arXiv: 1804.06437.

[18] Zachary C. Lipton. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, June 2018.

[19] Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. Gender Bias in Neural Natural Language Processing. *arXiv:1807.11714 [cs]*, May 2019. arXiv: 1807.11714.

[20] Aman Madaan, Amrith Setlur, Tanmay Parekh, Barnabas Poczos, Graham Neubig, Yiming Yang, Ruslan Salakhutdinov, Alan W. Black, and Shrimai Prabhumoye. Politeness Transfer: A Tag and Generate Approach. *arXiv:2004.14257 [cs]*, May 2020. arXiv: 2004.14257.

[21] Jianmo Ni, Jiacheng Li, and Julian McAuley. 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, Hong Kong, China, November 2019. Association for Computational Linguistics.

[22] Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. Right for the Right Reasons: Training Differentiable Models by Constraining their Explanations. *arXiv:1703.03717 [cs, stat]*, May 2017. arXiv: 1703.03717.

[23] Axel Sauer and Andreas Geiger. Counterfactual Generative Networks. *arXiv:2101.06046 [cs]*, January 2021. arXiv: 2101.06046.

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. *arXiv:1706.03762 [cs]*, December 2017. arXiv: 1706.03762.

[25] Hongning Wang, Yue Lu, and Chengxiang Zhai. Latent aspect rating analysis on review text data: a rating regression approach. In *In Proceedings of KDD '10*, pages 783–792. ACM, 2010.

[26] R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 2004.

[27] Xing Wu, Tao Zhang, Liangjun Zang, Jizhong Han, and Songlin Hu. "Mask and Infill" : Applying Masked Language Model to Sentiment Transfer. *arXiv:1908.08039 [cs]*, August 2019. arXiv: 1908.08039.

[28] Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient. *arXiv:1609.05473 [cs]*, August 2017. arXiv: 1609.05473.

[29] Mo Yu, Shiyu Chang, Yang Zhang, and Tommi Jaakkola. Rethinking Cooperative Rationalization: Introspective Extraction and Complement Control. 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 4092–4101, Hong Kong, China, 2019. Association for Computational Linguistics.

[30] Xiangji Zeng, Yunliang Li, Yuchen Zhai, and Yin Zhang. Counterfactual Generator: A Weakly-Supervised Method for Named Entity Recognition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7270–7280, Online, November 2020. Association for Computational Linguistics.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section **??**.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]

(d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] in supplemental material
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] in supplemental material
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] We followed convention for these datasets and for rationale models generally in not varying the random seed. See [16], [5], [6]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] in supplemental material
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] in supplemental material
   (b) Did you mention the license of the assets? [Yes] in supplemental material
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] in supplemental material. New, highly-correlated split on RateBeer data.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Appendix

Optionally include extra information (complete proofs, additional experiments and plots) in the appendix. This section will often be part of the supplemental material.