A Novel Estimator of Mutual Information for Learning to Disentangle Textual Representations

Pierre Colombo, Chloé Clavel, Pablo Piantanida

A Novel Estimator of Mutual Information for Learning to Disentangle Textual Representations

Pierre Colombo[†⋆], **Pablo Piantanida**[⋆], **Chloé Clavel**[⋆]
[⋆]Télécom ParisTech, Université Paris Saclay
† IBM GBS France
[⋆]Laboratoire des Signaux et Systèmes (L2S), CentraleSupelec CNRS Universite Paris-Saclay
pierre.colombo@ibm.com
chloe.clavel@telecom-paris.fr
pablo.piantanida@centralesupelec.fr

Abstract

Learning disentangled representations of textual data is essential for many natural language tasks such as fair classification, style transfer and sentence generation, among others. The existent dominant approaches in the context of text data either rely on training an adversary (discriminator) that aims at making attribute values difficult to be inferred from the latent code or rely on minimising variational bounds of the mutual information between latent code and the value attribute. However, the available methods suffer of the impossibility to provide a fine-grained control of the degree (or force) of disentanglement. In contrast to adversarial methods, which are remarkably simple, although the adversary seems to be performing perfectly well during the training phase, after it is completed a fair amount of information about the undesired attribute still remains. This paper introduces a novel variational upper bound to the mutual information between an attribute and the latent code of an encoder. Our bound aims at controlling the approximation error via the Renyi's divergence, leading to both better disentangled representations and in particular, a precise control of the desirable degree of disentanglement than state-of-the-art methods proposed for textual data. Furthermore, it does not suffer from the degeneracy of other losses in multi-class scenarios. We show the superiority of this method on fair classification and on textual style transfer tasks. Additionally, we provide new insights illustrating various trade-offs in style transfer when attempting to learn disentangled representations and quality of the generated sentence.

1 Introduction

Learning disentangled representations hold a central place to build rich embeddings of high-dimensional data. For a representation to be disentangled implies that it factorizes some latent cause or causes of variation as formulated by (Bengio et al., 2013). For example, if there are two causes for the transformations in the data that do not generally happen together and are statistically distinguishable (e.g., factors occur independently), a maximally disentangled representation is expected to present a sparse structure that separates those causes. Disentangled representations have been shown to be useful for a large variety of data, such as video (Hsieh et al., 2018), image (Sanchez et al., 2019), text (John et al., 2018), audio (Hung et al., 2018), among others, and applied to many different tasks, *e.g.*, robust and fair classification (Elazar and Goldberg, 2018), visual reasoning (van Steenkiste et al., 2019), style transfer (Fu et al., 2017), conditional generation (Denton et al., 2017; Burgess et al., 2018), few shot learning (Kumar Verma et al., 2018), among others.

In this work, we focus our attention on learning disentangled representations for text, as it remains overlooked by (John et al., 2018). Perhaps, one of the most popular applications of disentanglement in textual data is fair classification (Elazar and Goldberg, 2018; Barrett et al., 2019) and sentence generation tasks such as style transfer (John et al., 2018) or conditional sentence generation (Cheng et al., 2020b). For fair classification, perfectly disentangled latent representations can be used to ensure fairness as the decisions are taken based on representations which are statistically independent from–or at least carrying limited information about–the protected attributes. However, there exists a trade-offs between full disentangled representations and performances on the target task, as shown by (Feutry et al., 2018), among others. For sequence generation and in particular, for style transfer, learning disentangled representations aim at allowing an easier transfer of the desired style. To the best of our knowledge, a depth study of the relationship between disentangled representa-

tions based either on adversarial losses solely or on $vCLUB - S$ and quality of the generated sentences remains overlooked. Most of the previous studies have been focusing on either trade-offs between metrics computed on the generated sentences (Tikhonov et al., 2019) or performance evaluation of the disentanglement as part of (or convoluted with) more complex modules. This enhances the need to provide a fair evaluation of disentanglement methods by isolating their individual contributions (Yamshchikov et al., 2019; Cheng et al., 2020b). Methods to enforce disentangled representations can be grouped into two different categories. The first category relies on an adversarial term in the training objective that aims at ensuring that sensitive attribute values (*e.g.* race, sex, style) as statistically independent as possible from the encoded latent representation. Interestingly enough, several works (John et al., 2018; Elazar and Goldberg, 2018; Bao et al., 2019; Yi et al., 2020; Jain et al., 2019; Zhang et al., 2018; Hu et al., 2017), Elazar and Goldberg (2018) have recently shown that even though the adversary teacher seems to be performing remarkably well during training, after the training phase, a fair amount of information about the sensitive attributes still remains, and can be extracted from the encoded representation. The second category aim at minimising Mutual Information (MI) between encoded latent representation and the sensitive attribute values, *i.e.*, without resorting to an adversarial discriminator. MI acts as an universal measure of dependence since it captures non-linear and statistical dependencies of high orders between the involved quantities (Kinney and Atwal, 2014). However, estimating MI has been a long-standing challenge, in particular when dealing with high-dimensional data (Paninski, 2003; Pichler et al., 2020). Recent methods rely on variational upper bounds. For instance, (Cheng et al., 2020b) study `vCLUB-S` (Cheng et al., 2020a) for sentence generation tasks. Although this approach improves on previous state-of-the-art methods, it does not allow to fine-tuning of the desired degree of disentanglement, i.e., it enforces light or strong levels of disentanglement where only few features relevant to the input sentence remain (see Feutry et al. (2018) for further discussion).

1.1 Our Contributions

We develop new tools to build disentangled textual representations and evaluate them on fair classification and two sentence generation tasks, namely, style transfer and conditional sentence generation. Our main contributions are summarized below:

- *A novel objective to train disentangled representations from attributes.* To overcome some of the limitations of both adversarial losses and `vCLUB-S` we derive a novel upper bound to the MI which aims at correcting the approximation error via either the Kullback-Leibler (Ali and Silvey, 1966) or Renyi (Rényi et al., 1961) divergences. This correction terms appears to be a key feature to fine-tuning the degree of disentanglement compared to `vCLUB-S`.

- *Applications and numerical results.* First, we demonstrate that the aforementioned surrogate is better suited than the widely used adversarial losses as well as `vCLUB-S` as it can provide better disentangled textual representations while allowing *fine-tuning of the desired degree of disentanglement*. In particular, we show that our method offers a better accuracy versus disentanglement trade-offs for fair classification tasks. We additionally demonstrate that our surrogate outperforms both methods when learning disentangled representations for style transfer and conditional sentence generation while not suffering (or degenerating) when the number of classes is greater than two, which is an apparent limitation of adversarial training. By isolating the disentanglement module, we identify and report existing trade-offs between different degree of disentanglement and quality of generated sentences. The later includes content preservation between input and generated sentences and accuracy on the generated style.

2 Main Definitions and Related Works

We introduce notations, tasks, and closely related work. Consider a training set $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ of n sentences $x_i \in \mathcal{X}$ paired with attribute values $y_i \in \mathcal{Y} \equiv \{1, \dots, |\mathcal{Y}|\}$ which indicates a discrete attribute to be disentangled from the resulting representations. We study the following scenarios:

Disentangled representations. Learning disentangled representations consists in learning a model $\mathcal{M} : \mathcal{X} \to \mathcal{R}^d$ that maps feature inputs X to a vector of dimension d that retains as much as possible information of the original content from the input

sentence but as little as possible about the undesired attribute Y. In this framework, content is defined as any relevant information present in X that does not depend on Y.

Applications to binary fair classification. The task of fair classification through disentangled representations aims at building representations that are independent of selective discrete (sensitive) attributes (*e.g.*, gender or race). This task consists in learning a model $\mathcal{M} : \mathcal{X} \rightarrow \{0, 1\}$ that maps any input x to a label $l \in \{0, 1\}$. The goal of the learner is to build a predictor that assigns each x to either 0 or 1 "oblivious" of the protected attribute y. Recently, much progress has been made on devising appropriate means of fairness, *e.g.*, (Zemel et al., 2013; Zafar et al., 2017; Mohri et al., 2019). In particular, (Xie et al., 2017; Barrett et al., 2019; Elazar and Goldberg, 2018) approach the problem based on adversarial losses. More precisely, these approaches consist in learning an encoder that maps x into a representation vector h_x, a critic C_{θ_c} which attempts to predict y, and an output classifier f_{θ_d} used to predict l based on the observed h_x. The classifier is said to be fair if there is no statistical information about y that is present in h_x (Xie et al., 2017; Elazar and Goldberg, 2018).

Applications to conditional sentence generation. The task of conditional sentence generation consists in taking an input text containing specific stylistic properties to then generate a realistic (synthetic) text containing potentially different stylistic properties. It requests to learn a model $\mathcal{M} : \mathcal{X} \times \mathcal{Y} \rightarrow \mathcal{X}$ that maps a pair of inputs (x, y^t) to a sentence x^g, where the outcome sentence should retain as much as possible of the original content from the input sentence while having (potentially a new) attribute y^g. Proposed approaches to tackle textual style transfer (Zhang et al., 2020; Xu et al., 2019) can be divided into two main categories. The first category (Prabhumoye et al., 2018; Lample et al., 2018) uses cycle losses based on back translation (Wieting et al., 2017) to ensure that the content is preserved during the transformation. Whereas, the second category look to explicitly separate attributes from the content. This constraint is enforced using either adversarial training (Fu et al., 2017; Hu et al., 2017; Zhang et al., 2018; Yamshchikov et al., 2019) or MI minimisation using `vCLUB-S` (Cheng et al., 2020b). Traditional adversarial training is based on an encoder that aims to fool the adversary discriminator by removing attribute information from the content embedding (Elazar and Goldberg, 2018). As we will observe, the more the representations are disentangled the easier is to transfer the style but at the same time the less the content is preserved. In order to approach the sequence generation tasks, we build on the Style-embedding Model by (John et al., 2018) (StyleEmb) which uses adversarial losses introduced in prior work for these dedicated tasks. During the training phase, the input sentence is fed to a sentence encoder, namely f_{θ_e}, while the input style is fed to a separated style encoder, namely $f^s_{\theta_e}$. During the inference phase, the desired style–potentially different from the input style–is provided as input along with the input sentence.

3 Model and Training Objective

This section describes the proposed approach to learn disentangled representations. We first review MI along with the model overview and then, we derive the variational bound we will use, and discuss connections with adversarial losses.

3.1 Model Overview

The MI is a key concept in information theory for measuring high-order statistical dependencies between random quantities. Given two random variables Z and Y, the MI is defined by

$$I(Z;Y) = \mathbb{E}_{ZY}\left[\log \frac{p_{ZY}(Z,Y)}{p_Z(Z)p_Y(Y)}\right], \quad (1)$$

where p_{ZY} is the joint probability density function (pdf) of the random variables (Z, Y), with p_Z and p_Y representing the respective marginal pdfs. MI is related to entropy $h(Y)$ and conditional entropy $h(Y|Z)$ as follows:

$$I(Z;Y) = h(Y) - h(Y|Z). \quad (2)$$

Our models for fair classification and sequence generation share a similar structure. These rely on an encoder that takes as input a random sentence X and maps it to a random representation Z using a deep encoder denoted by f_{θ_e}. Then, classification and sentence generation are performed using either a classifier or an auto-regressive decoder denoted by f_{θ_d}. We aim at minimizing MI between the latent code represented by the Random Variable (RV) $Z = f_{\theta_e}(X)$ and the desired attribute represented by the RV Y. The objective of interest $\mathcal{L}(f_{\theta_e})$ is defined as:

$$\mathcal{L}(f_{\theta_e}) \equiv \underbrace{\mathcal{L}_{down.}(f_{\theta_e})}_{\text{downstream task}} + \lambda \cdot \underbrace{I(f_{\theta_e}(X);Y)}_{\text{disentangled}}, \quad (3)$$

where $\mathcal{L}_{down.}$ represents a downstream specific (target task) loss and λ is a meta-parameter that controls the sensitive trade-off between disentanglement (*i.e.*, minimizing MI) and success in the downstream task (*i.e.*, minimizing the target loss). In Sec. 5, we illustrate theses different trade-offs.

Applications to fair classification and sentence generation. For fair classification, we follow standard practices and optimize the cross-entropy between prediction and ground-truth labels. In the sentence generation task $\mathcal{L}_{down.}$ represents the negative log-likelihood between individual tokens.

3.2 A Novel Upper Bound on MI

Estimating the MI is a long-standing challenge as the exact computation (Paninski, 2003) is only tractable for discrete variables, or for a limited family of problems where the underlying data-distribution satisfies smoothing properties, see recent work by (Pichler et al., 2020). Different from previous approaches leading to variational lower bounds (Belghazi et al., 2018; Hjelm et al., 2018; Oord et al., 2018), in this paper we derive an estimator based on a variational upper bound to the MI which control the approximation error based on the Kullback-Leibler and the Renyi divergences (Daudel et al., 2020).

Theorem 1 *(Variational upper bound on MI) Let (Z, Y) be an arbitrary pair of RVs with $(Z, Y) \sim p_{ZY}$ according to some underlying pdf, and let $q_{\widehat{Y}|Z}$ be a conditional variational distribution on the attributes satisfying $p_{ZY} \ll p_Z \cdot q_{\widehat{Y}|Z}$, i.e., absolutely continuous. Then, we have that*

$$I(Z;Y) \leq \mathbb{E}_Y \left[-\log \int q_{\widehat{Y}|Z}(Y|z) p_Z(z) dz \right] + \mathbb{E}_{YZ} \left[\log q_{\widehat{Y}|Z}(Y|Z) \right] + KL\big(p_{ZY} \| p_Z \cdot q_{\widehat{Y}|Z}\big), \tag{4}$$

where $KL\big(p_{ZY} \| p_Z \cdot q_{\widehat{Y}|Z}\big)$ denotes the KL divergence. Similarly, we have that for any $\alpha > 1$,

$$I(Z;Y) \leq \mathbb{E}_Y \left[-\log \int q_{\widehat{Y}|Z}(Y|z) p_Z(z) dz \right] + \mathbb{E}_{YZ} \left[\log q_{\widehat{Y}|Z}(Y|Z) \right] + D_\alpha\big(p_{ZY} \| p_Z \cdot q_{\widehat{Y}|Z}\big), \tag{5}$$

where $(\alpha - 1) D_\alpha\big(p_{ZY} \| p_Z \cdot q_{\widehat{Y}|Z}\big) = \log \mathbb{E}_{ZY}[R^{\alpha-1}(Z,Y)]$ denotes the Renyi divergence and $R(z,y) = \frac{p_{Y|Z}(y|z)}{q_{\widehat{Y}|Z}(y|z)}$, for $(z,y) \in Supp(p_{ZY})$.

Proof: The upper bound on $H(Y)$ is a direct application of the the (Donsker and Varadhan, 1985) representation of KL divergence while the lower bound on $H(Y|Z)$ follows from the monotonicity property of the function: $\alpha \mapsto D_\alpha\big(p_{ZY} \| p_Z \cdot q_{\widehat{Y}|Z}\big)$. Further details are relegated to Appendix A.

Remark: It is worth to emphasise that the KL divergence in (4) and Renyi divergence in (5) control the approximation error between the exact entropy and its corresponding bound.

From theoretical bounds to trainable surrogates to minimize MI: It is easy to check that the inequalities in (Eq. 4) and (Eq. 5) are tight provided that $p_{ZY} \equiv p_Z \cdot q_{\widehat{Y}|Z}$ *almost surely* for some adequate choice of the variational distribution. However, the evaluation of these bounds requires to obtain an estimate of the density-ratio $R(z,y)$. Density-ratio estimation has been widely studied in the literature (see (Sugiyama et al., 2012) and references therein) and confidence bounds has been reported by (Kpotufe, 2017) under some smoothing assumption on underlying data-distribution p_{ZY}. In this work, we will estimate this ratio by using a critic C_{θ_R} which is trained to differentiate between a balanced dataset of positive i.i.d samples coming from p_{ZY} and negative i.i.d samples coming from $q_{\widehat{Y}|Z} \cdot p_Z$. Then, for any pair (z,y), the density-ratio can be estimated by $R(z,y) \approx \frac{\sigma(C_{\theta_R}(z,y))}{1-\sigma(C_{\theta_R}(z,y))}$, where $\sigma(\cdot)$ indicates the sigmoid function and $C_{\theta_R}(z,y)$ is the unnormalized output of the critic. It is worth to mention that after estimating this ratio, the previous upper bounds may not be strict bounds so we will refer them as surrogates.

3.3 Comparison to existing methods

Adversarial approaches: In order to enhance our understanding of why the proposed approach based on the minimization of the MI using our variational upper bound in Th. 1 may lead to a better training objective than previous adversarial losses, we discuss below the explicit relationship between MI and cross-entropy loss. Let $Y \in \mathcal{Y}$ denote a random attribute and let Z be a possibly high-dimensional representation that needs to be disentangled from Y. Then,

$$\begin{aligned} I(Z;Y) &\geq H(Y) - \mathbb{E}_{YZ} \left[\log q_{\widehat{Y}|Z}(Y|Z) \right] \\ &= \text{Const} - \text{CE}(\widehat{Y}|Z), \end{aligned} \tag{6}$$

where $\text{CE}(\widehat{Y}|Z)$ denotes the cross-entropy corresponding to the adversarial discriminator $q_{\widehat{Y}|Z}$, not-

ing that Y comes from an unknown distribution on which we have no influence $H(Y)$ is an unknown constant, and using that the approximation error: $\mathrm{KL}\big(q_{ZY}\|q_{\widehat{Y}|Z}\cdot p_Z\big) = \mathrm{CE}(\widehat{Y}|Z) - H(Y|Z)$. Eq. 6 shows that the cross-entropy loss leads to a lower bound (up to a constant) on the MI. Although the cross-entropy can lead to good estimates of the conditional entropy, the adversarial approaches for classification and sequence generation by (Barrett et al., 2019; John et al., 2018) which consists in maximizing the cross-entropy, induces a degeneracy (unbounded loss) as λ increases in the underlying optimization problem. As we will observe in next section, our variational upper bound in Th. 1 can overcome this issue, in particular for $|\mathcal{Y}| > 2$.

vCLUB-S: Different from our method, Cheng et al. (2020a) introduce I_{vCLUB} which is an upper bound on MI defined by

$$\begin{aligned} I_{\text{vCLUB}}(Y;Z) =& \mathbb{E}_{YZ}[\log p_{Y|Z}(Y|Z)] \\ & - \mathbb{E}_Y\mathbb{E}_Z[\log p_{Y|Z}(Y|Z)]. \end{aligned} \tag{7}$$

It would be worth to mention that this bound follows a similar approach to the previously introduced bound in (Feutry et al., 2018).

4 Experimental Setting

4.1 Datasets

Fair classification task. We follow the experimental protocol of (Elazar and Goldberg, 2018). The main task consists in predicting a binary label representing either the sentiment (positive/negative) or the mention. The mention task aims at predicting if a tweet is conversational. Here the considered protected attribute is the race. The dataset has been automatically constructed from DIAL corpus (Blodgett et al., 2016) which contained race annotations over 50 Million of tweets. Sentiment tweets are extracted using a list of predefined emojis and mentions are identified using @mentions tokens. The final dataset contains 160k tweets for the training and two splits of 10K tweets for validation and testing. Splits are balanced such that the random estimator is likely to achieve 50% accuracy.
Style Transfer For our sentence generation task, we conduct experiments on three different datasets extracted from restaurant reviews in Yelp. The first dataset, referred to as SYelp, contains 444101, 63483, and 126670 labelled short reviews (at most 20 words) for train, validation, and test, respectively. For each review a binary label is assigned depending on its polarity. Following (Lample et al., 2018), we use a second version of Yelp, referred to as FYelp, with longer reviews (at most 70 words). It contains five coarse-grained restaurant category labels (*e.g.*, Asian, American, Mexican, Bars and Dessert). The multi-category FYelp is used to access the generalization capabilities of our methods to a multi-class scenario.

4.2 Metrics for Performance Evaluation

Efficiency measure of the disentanglement methods. (Barrett et al., 2019) report that offline classifiers (post training) outperform clearly adversarial discriminators. We will re-training a classifier on the latent representation learnt by the model and we will report its accuracy.

Measure of performance within the fair classification task. In the fair classification task we aim at maximizing accuracy on the target task and so we will report the corresponding accuracy.

Measure of performance within sentence generation tasks. Sentences generated by the model are expected to be fluent, to preserve the input content and to contain the desired style. For style transfer, the desired style is different from the input style while for conditional sentence generation, both input and output styles should be similar. Nevertheless, automatic evaluation of generative models for text is still an open problem. We measure the style of the output sentence by using a fastText classifier (Joulin et al., 2016b). For content preservation, we follow (John et al., 2018) and compute both: (i) the cosine measure between source and generated sentence embeddings, which are the concatenation of min, max, and mean of word embedding (sentiment words removed), and (ii) the BLEU score between generated text and the input using SACREBLEU from (Post, 2018). Motivated by previous work, we evaluate the fluency of the language with the perplexity given by a GPT-2 (Radford et al., 2019) pretrained model performing fine-tuning on the training corpus. We choose to report the log-perplexity since we believe it can better reflects the uncertainty of the language model (a small variation in the model loss would induce a large change in the perplexity due to the exponential term). Besides the automatic evaluation, we further test our disentangled representation effectiveness by human evaluation results are presented in Tab. 1.
Conventions and abbreviations. *Adv* refers to a model trained using the adversarial loss;

vCLUB-S, KL refers to a model trained using the vCLUB-S and KL surrogate (see Eq. 14) respectively; and D_α refers to a model trained based on the α-Renyi surrogate (Eq. 15), for $\alpha \in \{1.3, 1.5, 1.8\}$.

5 Numerical Results

In this section, we present our results on the fair classification and binary sequence generation tasks, see Ssec. 5.1 and Ssec. 5.2, respectively. We additionally show that our variational surrogates to the MI–contrarily to adversarial losses–do not suffer in multi-class scenarios (see Ssec. 5.3).

5.1 Applications to Fairness

Upper bound on performances. We first examine how much of the protected attribute we can be recovered from an unfair classifier (*i.e.*, trained without adversarial loss) and how well does such classifier perform. Results are reported in Fig. 1. We observe that we achieve similar scores than the ones reported in previous studies (Barrett et al., 2019; Elazar and Goldberg, 2018). This experiment shows that, when training to solve the main task, the classifier learns information about the protected attribute, *i.e.*, the attacker's accuracy is better than random guessing. In the following, we compare the different proposed methods to disentangle representations and obtain a fairer classifier.

Methods comparisons. Fig. 1 shows the results of the different models and illustrates the trade-offs between disentangled representations and the target task accuracy. Results are reported on the testset for both sentiment and mention tasks when race is the protected. We observe that the classifier trained with an adversarial loss degenerates for $\lambda > 5$ since the adversarial term in Eq. 3 is influencing much the global gradient than the downstream term (*i.e.*, cross-entropy loss between predicted and golden distribution). Remarkably, both models trained to minimize either the KL or the Renyi surrogate do not suffer much from the aforementioned multi-class problem. For both tasks, we observe that the KL and the Renyi surrogates can offer better disentangled representations than those induced by adversarial approaches. In this task, both the KL and Renyi achieve perfect disentangled representations (*i.e.*, random guessing accuracy on protected attributes) with a 5% drop in the accuracy of the target task, when perfectly masking the protected attributes. As a matter of fact, we observe that vCLUB-S provides only two regimes: either a "light" protection (attacker accuracy around 60%), with almost no loss in task accuracy ($\lambda < 1$), or a strong protection (attacker accuracy around 50%), where a few features relevant to the target task remain.[1] On the sentiment task, we can draw similar conclusions. However, the Renyi's surrogate achieves slightly better-disentangled representations. Overall, we can observe that our proposed surrogate enables good control of the degree of disentangling. Additionally, we do not observe a degenerated behaviour–as it is the case with adversarial losses–when λ increases. Furthermore, our surrogate allows simultaneously better disentangled representations while preserving the accuracy of the target task.

5.2 Applications to binary polarity transfer

In the previous section, we have shown that the proposed surrogates do not suffer from limitations of adversarial losses and allow to achieve better disentangled representations than existing methods relying on vCLUB-S. Disentanglement modules are a core block for a large number of both style transfer and conditional sentence generation algorithms (Tikhonov et al., 2019; Yamshchikov et al., 2019; Fu et al., 2017) that place explicit constraints to force disentangled representations. First, we assess the disentanglement quality and the control over desired level of disentanglement while changing the downstream term, which for the sentence generation task is the cross-entropy loss on individual token. Then, we exhibit the existing trade-offs between quality of generated sentences, measured by the metric introduced in Ssec. 4.2, and the resulting degree of disentanglement. The results are presented for SYelp

5.2.1 Evaluating disentanglement

Fig. 2a shows the adversary accuracy of the different methods as a function of λ. Similarly to the fair classification task, a fair amount of information can be recovered from the embedding learnt with adversarial loss. In addition, we observe a clear degradation of its performance for values $\lambda > 1$. In this setting, the Renyi surrogates achieves consistently better results in terms of disentanglement than the one minimizing the KL surrogate. The curve for Renyi's surrogates shows that exploring different values of λ allows good control of the

[1] This phenomenon is also reported in (Feutry et al., 2018) on a picture anonymization task.

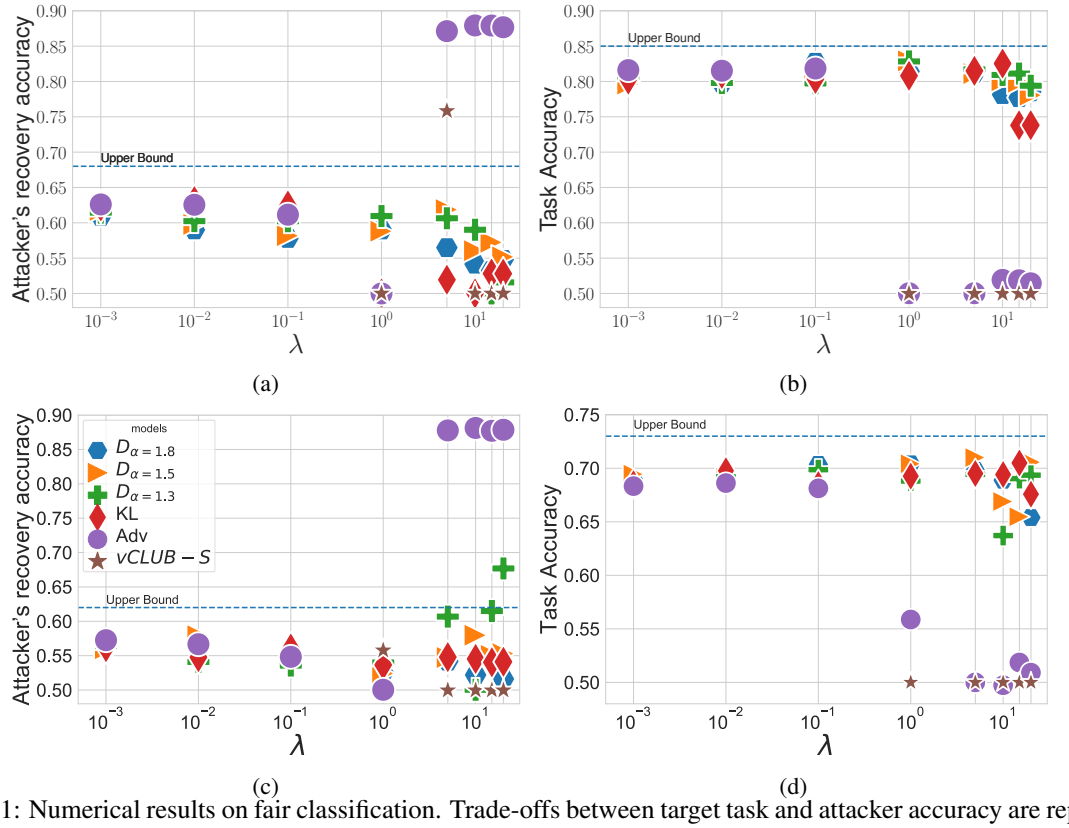


Figure 1: Numerical results on fair classification. Trade-offs between target task and attacker accuracy are reported in Fig. 1a, Fig. 1b for mention task, and Fig. 1c, Fig. 1d for sentiment task. For low values of λ some points coincide. As λ increases the level of disentanglement increases and the proposed methods using both KL (KL) and Reny divergences ($\mathcal{D}_\alpha$) clearly offer better control than existing methods.

disentanglement degree. Renyi surrogate generalizes well for sentence generation. Similarly to the fairness task vCLUB-S only offers two regimes: "light" disentanglement with very little polarity transfer and "strong" disentanglement.

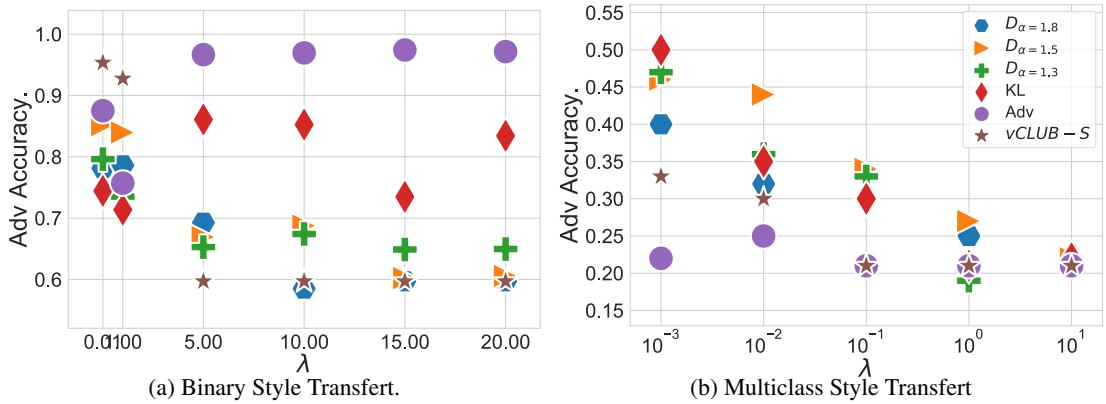
5.2.2 Disentanglement in Polarity Transfer

The quality of generated sentences are evaluated using the fluency (see Fig. 3c), the content preservation (see Fig. 3a), additional results using a cosine similarity are given in Appendix D, and polarity accuracy (see Fig. 3b). For style transfer, and for all models, we observe trade-offs between disentanglement and content preservation (measured by BLEU) and between fluency and disentanglement. Learning disentangled representations leads to poorer content preservation. As a matter of fact, similar conclusions can be drawn while measuring content with the cosine similarity (see Appendix D). For polarity accuracy, in non-degenerated cases (see below), we observe that the model is able to better transfer the sentiment in presence of disentangled representations. *Transferring style is easier with disentangled representations, however there is no free lunch here since disentangling also removes important information about the content*. It is worth noting that even in the "strong" disentanglement regime `vCLUB-S` struggles to transfer the polarity (accuracy of 40% for $\lambda \in \{1, 2, 10, 15\}$) where other models reach 80%. It is worth noting that similar conclusions hold for two different sentence generation tasks: style transfer and conditional generation, which tends to validate the current line of work that formulates text generation as generic text-to-text (Raffel et al., 2019).

Quality of generated sentences. Examples of generated sentences are given in Tab. 2 , providing qualitative examples that illustrate the previously observed trade-offs. The adversarial loss degenerates for values $\lambda \geq 5$ and a stuttering phenomenon appears (Holtzman et al., 2019). Tab. 1 gathers results of human evaluation and show that our surrogates can better disentangle style while preserving more content than available methods.

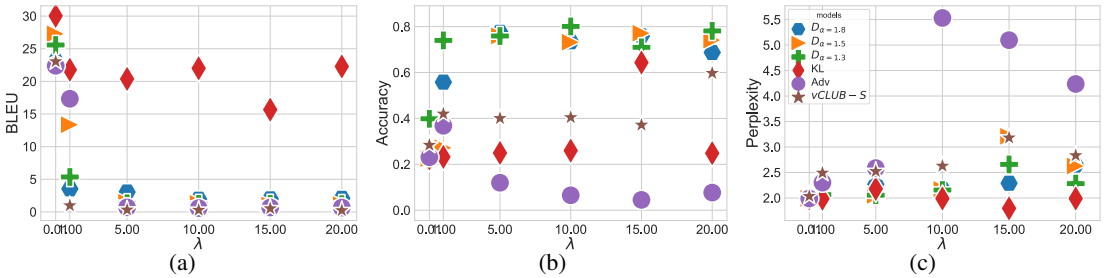
5.3 Adversarial Loss Fails to Disentangle when $|\mathcal{Y}| \geq 3$

In Fig. 2b we report the adversary accuracy of our different methods for the values of λ using FYelp



(a) Binary Style Transfert. (b) Multiclass Style Transfert

Figure 2: Disentanglement of representation learnt by f_{θ_e} in the binary (left) and multi-class (*i.e.*, $|\mathcal{Y}| = 5$) (right) sentence generation scenario. In the multi-class scenario the Adv degenerates for $\lambda \geq 0.01$ and offer no fined-grained control over the degree of disentanglement.



(a) (b) (c)

Figure 3: Numerical experiments on binary style transfer. Quality of generated sentences are evaluated using BLEU (Fig. 3a); style transfer accuracy (Fig. 3a); sentence fluency (Fig. 3c). We report existing trade-offs between disentanglement and sentence generation quality. Human evaluation is reported in Tab. 1.

dataset with category label. In the binary setting for $\lambda \leq 1$, models using adversarial loss can learn disentangled representations while in the multi-class setting, the adversarial loss degenerates for small values of λ (*i.e* sentences are no longer fluent as shown by the increase in perplexity in Fig. 4c). Minimizing MI based on our surrogates seems to mitigate the problem and offer a better control of the disentanglement degree for various values of λ than $vCLUB - S$. Further results are gathered in Appendix G.

6 Summary and Concluding Remarks

We devised a new alternative method to adversarial losses capable of learning disentangled textual representation. Our method does not require adversarial training and hence, it does not suffer in presence of multi-class setups. A key feature of this method is to account for the approximation error incurred when bounding the mutual information. Experiments show better trade-offs than both adversarial training and `vCLUB-S` on two fair classification tasks and demonstrate the efficiency to learn disentangled representations for sequence generation. As a matter of fact, there is no free-lunch for sentence generation tasks: *although transferring style is easier with disentangled representations, it also removes important information about the content*. The proposed method can replace the adversary in any kind of algorithms (Tikhonov et al., 2019; Fu et al., 2017) with no modifications. Future work includes testing with other type of labels such as dialog act (Chapuis et al., 2020; Colombo et al., 2020), emotions (Witon et al., 2018), opinion (Garcia et al., 2019) or speaker's stance and confidence (Dinkar et al., 2020). Since it allows more fine-grained control over the amount of disentanglement, we expect it to be easier to tune when combined with more complex models.

7 Acknowledgements

The authors would like to thanks Georg Pichler for the thorough reading. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 792464. The PhD of Pierre is fully founded by IBM GBS France in collaboration with Telecom Paris.

References

Syed Mumtaz Ali and Samuel D Silvey. 1966. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society: Series B (Methodological)*, 28(1):131–142.

Yu Bao, Hao Zhou, Shujian Huang, Lei Li, Lili Mou, Olga Vechtomova, Xinyu Dai, and Jiajun Chen. 2019. Generating sentences from disentangled syntactic and semantic spaces. *arXiv preprint arXiv:1907.05789*.

Maria Barrett, Yova Kementchedjhieva, Yanai Elazar, Desmond Elliott, and Anders Søgaard. 2019. Adversarial removal of demographic attributes revisited. 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 6331–6336.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. 2018. Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*.

Y. Bengio, A. Courville, and P. Vincent. 2013. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. 2018. Understanding disentangling in β-vae. *arXiv preprint arXiv:1804.03599*.

Emile Chapuis, Pierre Colombo, Matteo Manica, Matthieu Labeau, and Chloé Clavel. 2020. Hierarchical pre-training for sequence labelling in spoken dialog. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, EMNLP 2020, Online Event, 16-20 November 2020*, pages 2636–2648. Association for Computational Linguistics.

Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. 2020a. Club: A contrastive log-ratio upper bound of mutual information. In *International Conference on Machine Learning*, pages 1779–1788. PMLR.

Pengyu Cheng, Martin Renqiang Min, Dinghan Shen, Christopher Malon, Yizhe Zhang, Yitong Li, and Lawrence Carin. 2020b. Improving disentangled text representation learning with information-theoretic guidance. *arXiv preprint arXiv:2006.00693*.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Pierre Colombo, Emile Chapuis, Matteo Manica, Emmanuel Vignon, Giovanna Varni, and Chloe Clavel. 2020. Guiding attention in sequence-to-sequence models for dialogue act prediction. In *AAAI*, pages 7594–7601.

Pierre Colombo, Wojciech Witon, Ashutosh Modi, James Kennedy, and Mubbasir Kapadia. 2019. Affect-driven dialog generation. *arXiv preprint arXiv:1904.02793*.

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

Kamélia Daudel, Randal Douc, and François Portier. 2020. Infinite-dimensional gradient-based descent for alpha-divergence minimisation. Working paper or preprint.

Emily L Denton et al. 2017. Unsupervised learning of disentangled representations from video. In *Advances in neural information processing systems*, pages 4414–4423.

Tanvi Dinkar, Pierre Colombo, Matthieu Labeau, and Chloé Clavel. 2020. The importance of fillers for text representations of speech transcripts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 7985–7993. Association for Computational Linguistics.

MD Donsker and SRS Varadhan. 1985. Large deviations for stationary gaussian processes. *Communications in Mathematical Physics*, 97(1-2):187–210.

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. *arXiv preprint arXiv:1808.06640*.

Clément Feutry, Pablo Piantanida, Yoshua Bengio, and Pierre Duhamel. 2018. Learning anonymized representations with adversarial neural networks.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2017. Style transfer in text: Exploration and evaluation. *arXiv preprint arXiv:1711.06861*.

Alexandre Garcia, Pierre Colombo, Slim Essid, Florence d'Alché Buc, and Chloé Clavel. 2019. From

the token to the review: A hierarchical multimodal approach to opinion mining. *arXiv preprint arXiv:1908.11216*.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. 2018. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Jun-Ting Hsieh, Bingbin Liu, De-An Huang, Li F Fei-Fei, and Juan Carlos Niebles. 2018. Learning to decompose and disentangle representations for video prediction. In *Advances in Neural Information Processing Systems*, pages 517–526.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. *arXiv preprint arXiv:1703.00955*.

Yun-Ning Hung, Yi-An Chen, and Yi-Hsuan Yang. 2018. Learning disentangled representations for timber and pitch in music audio. *arXiv preprint arXiv:1811.03271*.

Parag Jain, Abhijit Mishra, Amar Prakash Azad, and Karthik Sankaranarayanan. 2019. Unsupervised controllable text formalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6554–6561.

Hamid Jalalzai, Pierre Colombo, Chloé Clavel, Eric Gaussier, Giovanna Varni, Emmanuel Vignon, and Anne Sabourin. 2020. Heavy-tailed representations, text polarity classification & data augmentation. *arXiv preprint arXiv:2003.11593*.

Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. 2018. Disentangled representation learning for non-parallel text style transfer. *arXiv preprint arXiv:1808.04339*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016a. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016b. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Justin B Kinney and Gurinder S Atwal. 2014. Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Sciences*, 111(9):3354–3359.

Samory Kpotufe. 2017. Lipschitz Density-Ratios, Structured Data, and Data-driven Tuning. volume 54 of *Proceedings of Machine Learning Research*, pages 1320–1328, Fort Lauderdale, FL, USA. PMLR.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. *arXiv preprint arXiv:1804.10959*.

Vinay Kumar Verma, Gundeep Arora, Ashish Mishra, and Piyush Rai. 2018. Generalized zero-shot learning via synthesized examples. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4281–4289.

Guillaume Lample, Sandeep Subramanian, Eric Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2018. Multiple-attribute text rewriting. In *International Conference on Learning Representations*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: A simple approach to sentiment and style transfer. *arXiv preprint arXiv:1804.06437*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. 2019. Agnostic federated learning. *arXiv preprint arXiv:1902.00146*.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Liam Paninski. 2003. Estimation of entropy and mutual information. *Neural computation*, 15(6):1191–1253.

Georg Pichler, Pablo Piantanida, and Günther Koliander. 2020. On the estimation of information measures of continuous distributions.

Matt Post. 2018. A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. *arXiv preprint arXiv:1804.09000*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Alfréd Rényi et al. 1961. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California.

Frank Rosenblatt. 1958. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386.

Eduardo Hugo Sanchez, Mathieu Serrurier, and Mathias Ortner. 2019. Learning disentangled representations via mutual information estimation. *arXiv preprint arXiv:1912.03915*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Sjoerd van Steenkiste, Francesco Locatello, Jürgen Schmidhuber, and Olivier Bachem. 2019. Are disentangled representations helpful for abstract visual reasoning? In *Advances in Neural Information Processing Systems*, pages 14245–14258.

Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. 2012. *Density Ratio Estimation in Machine Learning*, 1st edition. Cambridge University Press, USA.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Alexey Tikhonov, Viacheslav Shibaev, Aleksander Nagaev, Aigul Nugmanova, and Ivan P Yamshchikov. 2019. Style transfer for texts: Retrain, report errors, compare with rewrites. 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 3927–3936.

Tim Van Erven and Peter Harremos. 2014. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820.

John Wieting, Jonathan Mallinson, and Kevin Gimpel. 2017. Learning paraphrastic sentence embeddings from back-translated bitext. *arXiv preprint arXiv:1706.01847*.

Wojciech Witon, Pierre Colombo, Ashutosh Modi, and Mubbasir Kapadia. 2018. Disney at IEST 2018: Predicting emotions using an ensemble. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, WASSA@EMNLP 2018, Brussels, Belgium, October 31, 2018*, pages 248–253. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Qizhe Xie, Zihang Dai, Yulun Du, Eduard Hovy, and Graham Neubig. 2017. Controllable invariance through adversarial feature learning. In *Advances in Neural Information Processing Systems*, pages 585–596.

Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. 2015. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*.

Ruochen Xu, Tao Ge, and Furu Wei. 2019. Formality style transfer with hybrid textual annotations. *arXiv preprint arXiv:1903.06353*.

Ivan P Yamshchikov, Viacheslav Shibaev, Aleksander Nagaev, Jürgen Jost, and Alexey Tikhonov. 2019. Decomposing textual information for style transfer. *arXiv preprint arXiv:1909.12928*.

Xiaoyuan Yi, Zhenghao Liu, Wenhao Li, and Maosong Sun. 2020. Text style transfer via learning style instance supported latent space. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3801–3807. International Joint Conferences on Artificial Intelligence Organization.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pages 1171–1180.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. volume 28 of *Proceedings of Machine Learning Research*, pages 325–333, Atlanta, Georgia, USA. PMLR.

Ye Zhang, Nan Ding, and Radu Soricut. 2018. Shaped: Shared-private encoder-decoder for text style adaptation. *arXiv preprint arXiv:1804.04093*.

Yi Zhang, Tao Ge, and Xu Sun. 2020. Parallel data augmentation for formality style transfer. *arXiv preprint arXiv:2005.07522*.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.