Learning to Disentangle Textual Representations and Attributes via Mutual Information

Pierre Colombo,Chloe Clavel
S2A - Telecom Paris
{pierre.colombo,chloe.clavel}@telecom-paris.fr

Pablo Piantanida
Laboratoire des Signaux et Systemes, CentraleSupelec
CNRS, Universite Paris-Saclay,
pablo.piantanida@centralesupelec.fr

Abstract

Learning disentangled representations of textual data is essential for many natural language tasks such as fair classification (*e.g.* building classifiers whose decisions cannot disproportionately hurt or benefit specific groups identified by sensitive attributes), style transfer and sentence generation, among others. The existent dominant approaches in the context of text data have been based on training an adversary (discriminator or teacher) that aims at making attribute values difficult to be inferred from the latent code. Although these approaches are remarkably simple and even though the adversary seems to be performing perfectly during the training phase, after training is completed a fair amount of sensitive information to infer the attribute still remains. This paper investigates learning to disentangle representations by minimizing a novel variational (upper) bound of the mutual information between an identified attribute and the latent code of a deep neural network encoder. We demonstrate that our surrogate leads to better disentangled representations on both fair classification and sentence generation tasks while not suffering from the degeneracy of adversarial losses in multi-class scenarios. Furthermore, by optimizing the trade-off between the level of disentanglement and quality of the generated sentences for polarity transfer and sentence generation tasks, we provide some lights to the well-known debate on whether or not *"disentangled representations may be helpful for polarity transfer and sentence generation purposes"*.

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)), 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 & 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 & 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. (2020)). 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-off between full disentangled representations and performances on the target task, as shown by Feutry et al. (2018). On the other hand, for sequence generation and in particular for style transfer, the utility of learning disentangled representations remains an open question and only few studies focus on the effective impact of disentangling on the downstream (target task) performance.

The dominant approach to learn disentangled representations 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); Barrett et al. (2019); Elazar & Goldberg (2018); Bao et al. (2019); Yi et al. (2020); Jain et al. (2019); Zhang et al. (2018); Hu et al. (2017), Elazar & Goldberg (2018); Lample et al. (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.

1.1 Our Contributions

We study 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 can be summarized as follow:

- *A novel objective to train disentangled representations from attributes.* To overcome some of the limitations of adversarial losses to learn disentangled representations, we propose to minimize the Mutual Information (MI) between the latent code and the attribute, *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 & 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). Instead, we simple derive a novel surrogate (upper bound) to the MI based on Kullback-Leibler (Ali & Silvey (1966)) and Renyi (Rényi et al. (1961)) divergences.
- *Applications and numerical results.* First, we demonstrate that the aforementioned surrogate is better suited than the widely used adversarial losses 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, binary style transfer and conditional sentence generation. The later includes content preservation between input and generated sentences and a desired style. We believe that the present work is the first that provides an exhaustively study of the underlying trade-off that exists between disentanglement, style accuracy and content preservation in the context of sentence generation. Finally, we demonstrate that contrarily to adversarial losses, our method does not suffer (or degenerate) when the number of classes is greater than two, which is an apparent limitation of adversarial training.

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 $y_i \in \mathcal{Y} \equiv \{1, \ldots, |\mathcal{Y}|\}$ 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. Although variational methods have been widely used for learning disentangled representations for image, video or audio (He et al. (2020); Ansari & Soh (2019); Mathieu et al. (2019); Dupont (2018); Esmaeili et al. (2019); Kim & Mnih (2018); Shu et al. (2018); Hoffman et al. (2017); Kumar et al. (2017); Higgins et al. (2016)), textual embeddings are usually disentangled based on adversarial methods (Barrett et al. (2019)). In this work, we learn disentangled representation by minimizing the MI and thus, it is desirable to derive an upper (surrogate) bound to the MI. However, the available methods to estimate MI rely on variational lower bounds which are mostly used to maximize MI (*e.g.*, InfoMax Linsker (1988), and Belghazi et al. (2018); Hjelm et al. (2018); Oord et al. (2018), among others). As the main focus of this paper is not estimating MI, an extensive comparison with other MI estimators will be omitted (Kraskov et al. (2005); Poole et al. (2019)).

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 & 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 & Goldberg (2018)).

Applications to sentence generation. The task of 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 relies on adversarial training (Fu et al. (2017); Hu et al. (2017); Zhang et al. (2018)) to separate attributes from the content. Traditional training is based on an encoder that aims to fool the adversary discriminator by removing attribute information from the content embedding (Elazar & 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

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

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:

$$I(Z; Y) = h(Y) - h(Y|Z). \tag{2}$$

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 asw:

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

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 Variational Upper Bounds 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 and estimate a variational upper (surrogate) bound to MI which is 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*

$$I(Z;Y) \leq \mathbb{E}_Y \left[-\log \int_{R^d} Q_{\widehat{Y}|Z}(Y|z) P_Z(dz) \right] + \mathbb{E}_{YZ} \left[\log Q_{\widehat{Y}|Z}(Y|Z) \right] + KL\big(P_{ZY} \| P_Z Q_{\widehat{Y}|Z}\big), \tag{4}$$

where $KL\big(P_{ZY} \| P_Z Q_{\widehat{Y}|Z}\big)$ denotes the KL divergence. Similarly, we have for any $\alpha > 1$,

$$I(Z;Y) \leq \mathbb{E}_Y \left[-\log \int_{R^d} Q_{\widehat{Y}|Z}(Y|z) P_Z(dz) \right] + \mathbb{E}_{YZ} \left[\log Q_{\widehat{Y}|Z}(Y|Z) \right] + D_\alpha\big(P_{ZY} \| P_Z Q_{\widehat{Y}|Z}\big), \tag{5}$$

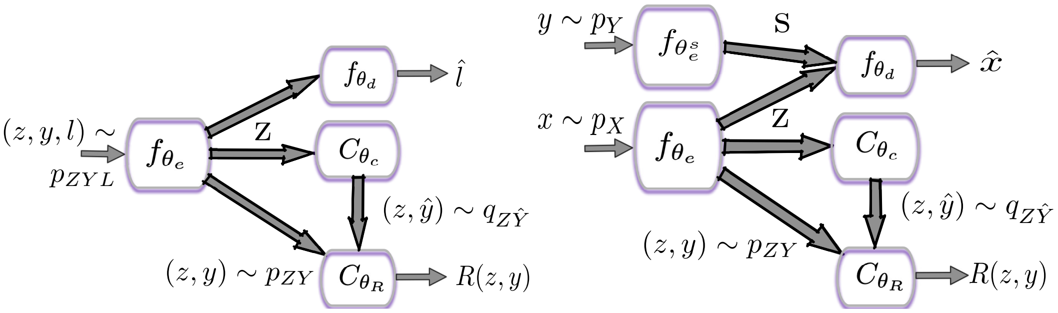
where $D_\alpha\big(P_{ZY} \| P_Z Q_{\widehat{Y}|Z}\big) = \frac{1}{\alpha-1} \log \mathbb{E}_{ZY}[R^{\alpha-1}(Z,Y)]$ denotes the Renyi divergence and $R(z,y) = \frac{P_{Y|Z}(y|z)}{Q_{\widehat{Y}|Z}(y|z)}$, for all pairs $(z,y) \in Supp(P_{ZY})$.

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

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. Overall a schema of the proposed methods is provided in Fig. 1.

3.3 Comparison to 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



(a) **Application to fair classification** (b) **Application to sentence generation**

Figure 1: Proposed methods based on Renyi's surrogate in Th. 1. f_{θ_e},$f^s_{\theta_e}$,C_{θ_c},C_{θ_R},R,f_{θ_d} respectively represents: the input sentence encoder; the style encoder (only used for sentence generation tasks); the attribute classifier; the classifier used to estimate R; the decoder which can be either a classifier (Fig. 1a) or a sequence decoder (Fig. 1b). Schemes of the baseline models are given in Fig. 7.

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,

$$I(Z;Y) \geq H(Y) - \mathbb{E}_{YZ}\left[\log Q_{\widehat{Y}|Z}(Y|Z)\right] \approx \log|\mathcal{Y}| - \mathrm{CE}(\widehat{Y}|Z), \tag{6}$$

where $\mathrm{CE}(\widehat{Y}|Z)$ denotes the cross-entropy corresponding to the adversarial discriminator $Q_{\widehat{Y}|Z}$, assuming Y is uniformly distributed and noting that $H(Y|Z) = \mathrm{CE}(\widehat{Y}|Z) - \mathrm{KL}\big(P_{ZY} \| Q_{\widehat{Y}|Z}P_Z\big)$. 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 $|\mathcal{Y}| > 2$.

4 Experimental Setting

4.1 Datasets

Fair classification task. We follow the experimental protocol of Elazar & 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.

Conditional sentence generation and polarity transfer tasks. 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. Comparizon to prior work can be found in Appendix D.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 both binary gender annotations (*i.e.*, annotated following Prabhumoye et al. (2018); Reddy & Knight (2016), results on this dataset are reported in Appendix E), and 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 scores.

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 one 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).

Conventions and abbreviations. Adv refers to a model trained using the adversarial loss; KL refers to a model trained using the KL surrogate, as described in Eq. 15; and D_α refers to a model trained based on the α-Renyi surrogate (Eq. 16), for $\alpha \in \{1.3, 1.5, 1.8\}$.

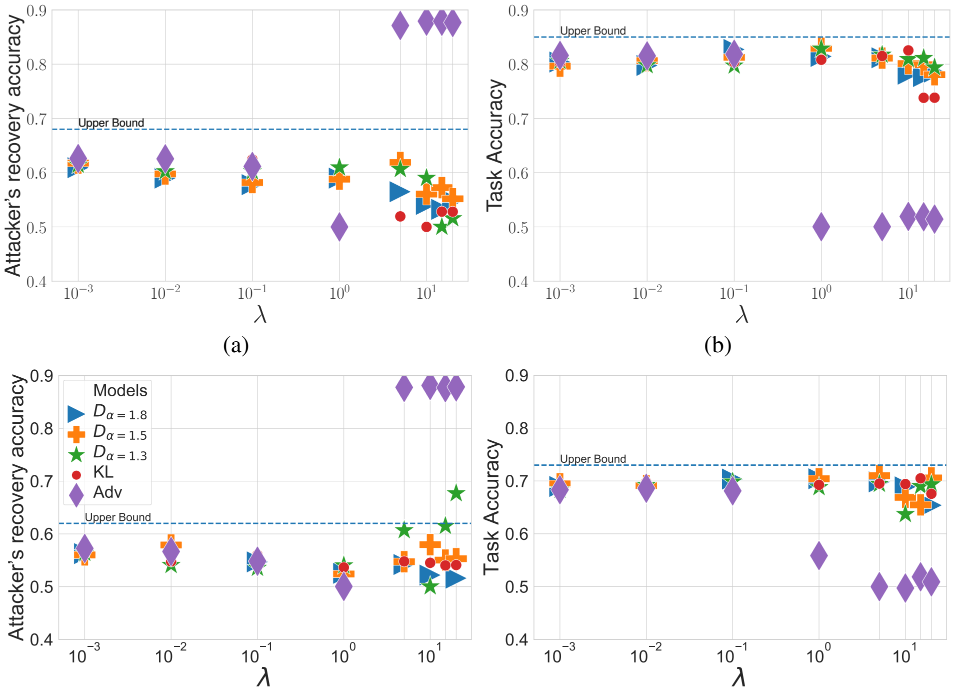
5 Numerical Results

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. 2. We observe that we achieve similar scores than the ones reported in previous studies Barrett et al. (2019); Elazar & 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. 2 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. 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 a better disentangled representations while better preserving the accuracy of the target task.



(a)

(b)

(c)

(d)

Figure 2: Numerical results on fair classification. Trade-offs between target task and attacker accuracy are reported in Fig. 2a, Fig. 2b for mention task, and Fig. 2c, Fig. 2d for sentiment task.

5.2 Applications to Sentence Generation with Binary Classes

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. Nevertheless, for sentence generation tasks, it remains an open question: *does the placement of explicit constraints to force disentangled representations lead to tackle style transfer and/or conditional sentence generation?* First, we assess the disentanglement quality 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. 3 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 disentanglement degree. Renyi surrogate generalizes well for sentence generation.

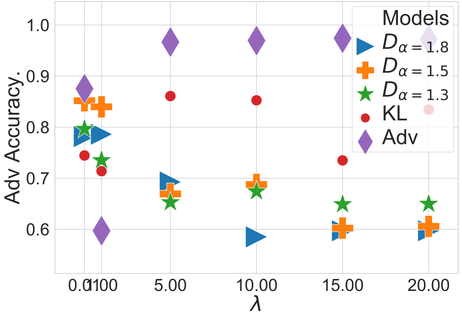
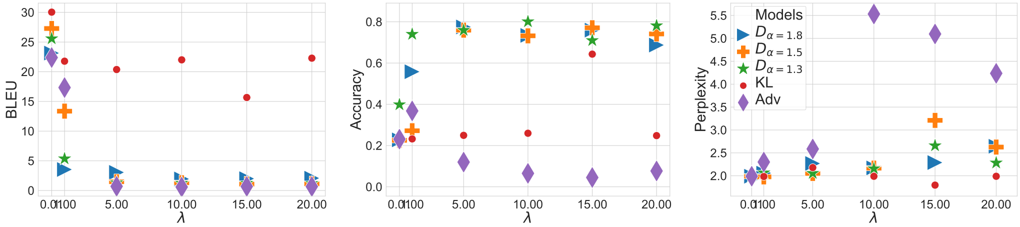


Figure 3: Disentanglement of the representations learnt by the encoder f_{θ_e} when the model is trained on a binary sentence generation task.

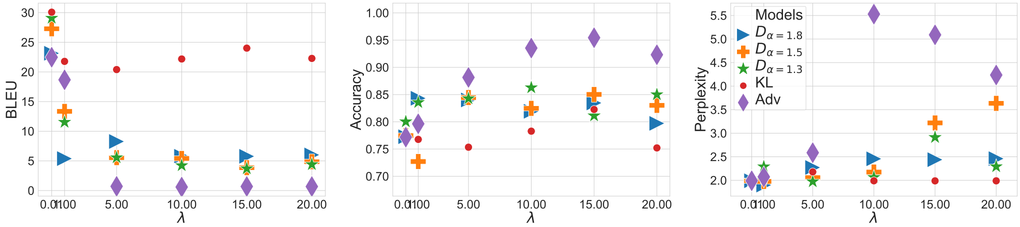
5.2.2 Disentanglement in Style Transfer and Conditional Sentence Generation

The quality of generated sentences are evaluated using the fluency (see Fig. 4c and Fig. 5c), the content preservation (see Fig. 4a and Fig. 5a), additional results using a cosinus similarity are



(a) (b) (c)

Figure 4: Numerical experiments on binary style transfer. Quality of generated sentences are evaluated using BLEU (Fig. 4a); style transfer accuracy (Fig. 4a); sentence fluency (Fig. 4c).



(a) (b) (c)

Figure 5: Numerical experiments on conditional sentence generation. Results include BLEU (Fig. 5a), style transfer accuracy (Fig. 5b) and sentence fluency (Fig. 5c).

given in Appendix D, and polarity accuracy (see Fig. 4b and Fig. 5b). For both style transfer and conditional sentence generation, 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 cosinus 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 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. 3 and 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).

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

In Fig. 6 we report the adversary accuracy of our different methods for the values of λ using FYelp 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. Minimizing MI based on our surrogates seems to mitigate the problem as we can observe a better control of the disentanglement degree for various values of λ. Further results are gathered in Appendix F.

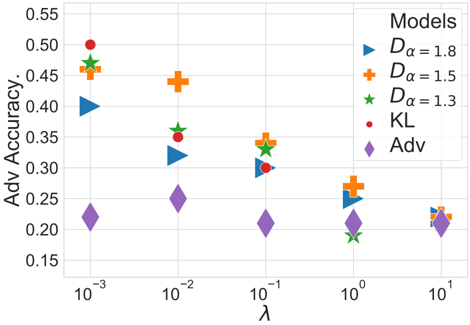


Figure 6: Disentanglement of representation learnt by f_{θ_e} in the multi-class sentence generation scenario (*i.e.*, $|\mathcal{Y}| = 5$).

6 Summary and Concluding Remarks

We devised a new method capable of learning disentangled textual representation. Different from most of existent approaches, our method does not require adversarial training and hence, it does not suffer in presence of multi-class setups. Experiments show better trade-offs on two fair classification tasks and demonstrate the efficiency of the method to control style for sentence generation tasks. 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 disentangled factors we studied are polarity (usually regarded as style, which is open to criticism) and demographic attributes corresponding to two applications: conditional sentence generation and fairness in classification, respectively. The former is important for the dialogue community because it enables the opportunity to better control the system's answer according to the user's inputs and profile. However, for an effective use in dialogue systems, future studies should focus on the development of new datasets annotated into labels which are more relevant to disentangle than polarity for dialog control (*e.g.*, formal, informal style, agent's emotion or personality) and to the development of dialogue-oriented evaluation methods such as perceptive test and user study.

References

Syed Mumtaz Ali and Samuel D Silvey. 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, 1966.

Abdul Fatir Ansari and Harold Soh. Hyperprior induced unsupervised disentanglement of latent representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3175–3182, 2019.

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

Maria Barrett, Yova Kementchedjhieva, Yanai Elazar, Desmond Elliott, and Anders Søgaard. 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)*, pp. 6331–6336, 2019.

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

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

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 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*, pp. 1119–1130, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1120. URL https://www.aclweb.org/anthology/D16-1120.

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

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

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

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

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

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

Kamélia Daudel, Randal Douc, and François Portier. Infinite-dimensional gradient-based descent for alpha-divergence minimisation. working paper or preprint, May 2020. URL https://hal.telecom-paris.fr/hal-02614605.

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

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

Emilien Dupont. Learning disentangled joint continuous and discrete representations. In *Advances in Neural Information Processing Systems*, pp. 710–720, 2018.

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

Babak Esmaeili, Hao Wu, Sarthak Jain, Alican Bozkurt, Narayanaswamy Siddharth, Brooks Paige, Dana H Brooks, Jennifer Dy, and Jan-Willem Meent. Structured disentangled representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2525–2534. PMLR, 2019.

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

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

Alexandre Garcia, Pierre Colombo, Slim Essid, Florence d'Alché Buc, and Chloé Clavel. From the token to the review: A hierarchical multimodal approach to opinion mining. *arXiv preprint arXiv:1908.11216*, 2019.

Junxian He, Xinyi Wang, Graham Neubig, and Taylor Berg-Kirkpatrick. A probabilistic formulation of unsupervised text style transfer. *arXiv preprint arXiv:2002.03912*, 2020.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. 2016.

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

Matthew D Hoffman, Carlos Riquelme, and Matthew J Johnson. The β-vae's implicit prior. In *Workshop on Bayesian Deep Learning, NIPS*, pp. 1–5, 2017.

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

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

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

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

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

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

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

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

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

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. *arXiv preprint arXiv:1802.05983*, 2018.

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

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

Samory Kpotufe. Lipschitz Density-Ratios, Structured Data, and Data-driven Tuning. volume 54 of *Proceedings of Machine Learning Research*, pp. 1320–1328, Fort Lauderdale, FL, USA, 20–22 Apr 2017. PMLR. URL `http://proceedings.mlr.press/v54/kpotufe17a.html`.

Alexander Kraskov, Harald Stögbauer, Ralph G Andrzejak, and Peter Grassberger. Hierarchical clustering using mutual information. *EPL (Europhysics Letters)*, 70(2):278, 2005.

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

Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. *arXiv preprint arXiv:1711.00848*, 2017.

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

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

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

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

Ralph Linsker. Self-organization in a perceptual network. *Computer*, 21(3):105–117, 1988.

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

Emile Mathieu, Tom Rainforth, N Siddharth, and Yee Whye Teh. Disentangling disentanglement in variational autoencoders. In *International Conference on Machine Learning*, pp. 4402–4412, 2019.

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

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

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

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

Ben Poole, Sherjil Ozair, Aaron van den Oord, Alexander A Alemi, and George Tucker. On variational bounds of mutual information. *arXiv preprint arXiv:1905.06922*, 2019.

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

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

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

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

Sravana Reddy and Kevin Knight. Obfuscating gender in social media writing. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pp. 17–26, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-5603. URL `https://www.aclweb.org/anthology/W16-5603`.

Alfréd Rényi et al. 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, 1961.

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

Rico Sennrich, Barry Haddow, and Alexandra Birch. 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)*, pp. 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1162. URL `https://www.aclweb.org/anthology/P16-1162`.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. Style transfer from non-parallel text by cross-alignment. In *Advances in neural information processing systems*, pp. 6830–6841, 2017.

Rui Shu, Shengjia Zhao, and Mykel J Kochenderfer. Rethinking style and content disentanglement in variational autoencoders. 2018.

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

Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density Ratio Estimation in Machine Learning*. Cambridge University Press, USA, 1st edition, 2012. ISBN 0521190177.

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

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

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

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

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. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771, 2019.

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

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

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

Xiaoyuan Yi, Zhenghao Liu, Wenhao Li, and Maosong Sun. Text style transfer via learning style instance supported latent space. In Christian Bessiere (ed.), *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pp. 3801–3807. International Joint Conferences on Artificial Intelligence Organization, 2020. URL https://doi.org/10.24963/ijcai.2020/526.

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

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. volume 28 of *Proceedings of Machine Learning Research*, pp. 325–333, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL http://proceedings.mlr.press/v28/zemel13.html.

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

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

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 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*, pp. 19–27, 2015.

A Additional Details On the Surrogates

A.1 Proof of Inequality Eq. 6

In this section, we provide a formal proof of the Eq. 6. 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 probability distribution on the discrete attributes satisfying $P_{ZY} \ll P_Z \cdot Q_{\widehat{Y}|Z}$, i.e., absolutely continuous.

$$I(Z;Y) \geq H(Y) - \mathrm{CE}(\hat{Y}|Z). \tag{7}$$

Proof:

$$\begin{aligned} I(Z;Y) &= H(Y) - H(Y|Z) && (8)\\ &= \log|\mathcal{Y}| - H(Y|Z), && (9)\end{aligned}$$

provided that Y is uniformly distributed.

We then need to find the relationship between the cross-entropy and the conditional entropy.

$$\begin{aligned} \mathrm{KL}(P_{YZ}\|Q_{\hat{Y}Z}) &= E_{YZ}\left[\log\frac{P_{Y|Z}(Y|Z)}{Q_{\hat{Y}|Z}(Y|Z)}\right] && (10)\\ &= E_{YZ}\big[\log P_{Y|Z}(Y|Z)\big] - E_{YZ}\big[\log Q_{\hat{Y}|Z}(Y|Z)\big] && (11)\\ &= -H(Y|Z) + \mathrm{CE}(\hat{Y}|Z). && (12)\end{aligned}$$

We know that $\mathrm{KL}(P_{YZ}\|Q_{\hat{Y}Z}) \geq 0$, thus $\mathrm{CE}(\hat{Y}|Z) \geq H(Y|Z)$ which gives the result.

The underlying hypothesis made by approximating the MI with an adversarial loss is that the contribution of gradient from $\mathrm{KL}(P_{YZ}\|Q_{\hat{Y}Z})$ to the bound is negligible.

A.2 Proof of Th. 1

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 probability distribution satisfying $P_{ZY} \ll P_Z \cdot Q_{\widehat{Y}|Z}$, i.e., absolutely continuous. To obtain an upper bound on the MI we need to upper bound the entropy $H(Y)$ and to lower bound the conditional entropy $H(Y|Z)$.

Upper bound on $H(Y)$. Since the KL divergence is non-negative, we have

$$\begin{aligned} H(Y) &\leq \mathbb{E}_Y\left[-\log Q_Y(Y)\right] && (13)\\ &= \mathbb{E}_Y\left[-\log\int_{R^d} Q_{\hat{Y}|Z}(Y|z)P_z(dz)\right]. && (14)\end{aligned}$$

Lower bounds on $H(Y|Z)$. We have the following inequalities:

$$H(Y|Z) = \mathbb{E}_{YZ}\left[-\log Q_{\hat{Y}|Z}(Y|Z)\right] - \mathrm{KL}(P_{YZ}\|P_Z Q_{\hat{Y}|Z}), \tag{15}$$

where $\mathrm{KL}(P_{YZ}\|P_Z Q_{\hat{Y}|Z})$ denotes the KL divergence. Furthermore, for arbitrary values $\alpha > 1$,

$$H(Y|Z) \leq \mathbb{E}_{YZ}\left[-\log Q_{\hat{Y}|Z}(Y|Z)\right] - D_\alpha(P_{YZ}\|P_Z Q_{\hat{Y}|Z}), \tag{16}$$

where

$$D_\alpha(P_{YZ}\|P_Z Q_{\hat{Y}|Z}) = \frac{1}{\alpha-1}\log\mathbb{E}_{ZY}\left[R^{\alpha-1}(Z,Y)\right]$$

is the Renyi divergence with

$$R(y,z) = \frac{P_{Y|Z}(y|z)}{Q_{\hat{Y}|Z}(y|z)}.$$

The proof of Eq. 15 is given in Ssec. A.1. In order to show Eq. 16, we remark that Renyi divergence is non-decreasing function $\alpha \mapsto D_\alpha(P_{ZY}\|P_Z Q_{\hat{Y}|Z})$ in $\alpha \in [0, +\infty)$ (the reader is refereed to Van Erven & Harremos (2014) for a detailed proof). Thus, we have

$$\mathrm{KL}(P_{ZY}\|P_Z Q_{\hat{Y}|Z}) \leq D_\alpha(P_{ZY}\|P_Z Q_{\hat{Y}|Z}), \qquad \forall\alpha > 1. \tag{17}$$

Therefore, from expression Eq. 15 we obtain the result.

A.3 Optimization of the Surrogates on MI

In this section, we give details to facilitate the practical implementation of our methods.

A.3.1 Computing the entropy $H(Y)$

$$\begin{aligned}\mathbb{E}_Y\left[-\log\int_{R^d} Q_{\hat{Y}|Z}(Y|z)P_Z(dz)\right] &\approx \mathbb{E}_Y\left[-\log\sum_{i=1}^{n} Q_{\hat{Y}|Z}(Y|z_i)\right] + \text{const.}\\ &\approx -\frac{1}{|\mathcal{Y}|}\sum_{j=1}^{|\mathcal{Y}|}\log\sum_{i=1}^{n} C_{\theta_c}(z_i)_{y_j} + \text{const.}\end{aligned} \tag{18}$$

where $C_{\theta_c}(z_i)_{y_j}$ is the y_j-th component of the normalised output of the classifier C_{θ_c}.

A.3.2 Computing the lower bound on $H(Y|Z)$

The upper bound hold for $\alpha > 1$,

$$H(Y|Z) \approx \text{CE}(Y|Z) - \hat{D}_\alpha(P_{ZY}\|P_Z Q_{\hat{Y}|Z}) \tag{19}$$

$$\approx -\frac{1}{n}\sum_{i=1}^{n}\log Q_{\hat{Y}|Z}(y_i|z_i) - \frac{1}{\alpha-1}\log\sum_{i=1}^{n} R^{\alpha-1}(z_i, y_i). \tag{20}$$

Estimating the density-ratio $R(z, y)$ In what follow we apply the so-called density-ratio trick to our specific setup. Suppose we have a balanced dataset of point $\{(y_i^p, z_i^p)\} \sim p_{YZ}$ and $\{(y_i^q, z_i^q)\} \sim Q_{\hat{Y}|Z}p_Z$ with $i \in [1, K]$. The density-ratio trick consists in training a classifier C_{θ_R} to distinguish between theses two distribution. Samples coming from p are labelled $u = 1$, samples coming from q are labelled $u = 0$. Thus, we can rewrite $R(z, y)$ as

$$R(z, y) = \frac{p_{Y|Z}(y, z)}{q_{\hat{Y}|Z}(y, z)} \tag{21}$$

$$= \frac{p_{YZ|U}(y, z|u = 0)}{p_{YZ|U}(y, z|u = 1)} \tag{22}$$

$$= \frac{P_{U|YZ}(u = 0|y, z)}{P_{U|YZ}(u = 1|y, z)}\frac{p_U(u = 1)}{p_U(u = 0)} \tag{23}$$

$$= \frac{P_{U|YZ}(u = 0|y, z)}{P_{U|YZ}(u = 1|y, z)} \tag{24}$$

$$= \frac{P_{U|YZ}(u = 0|y, z)}{1 - P_{U|YZ}(u = 0|y, z)}. \tag{25}$$

Obviously, the true posterior distribution $P_{U|YZ}$ is known. However, if C_{θ_R} is well trained, then $P_{U|YZ}(u = 0|y, z) \approx \sigma(C_{\theta_R}(y, z))$, where $\sigma(\cdot)$ denotes the sigmoid function. A detailled procedure for training is given in Algorithm 1.

B Additional Details on the Model

B.1 Baseline Schemas

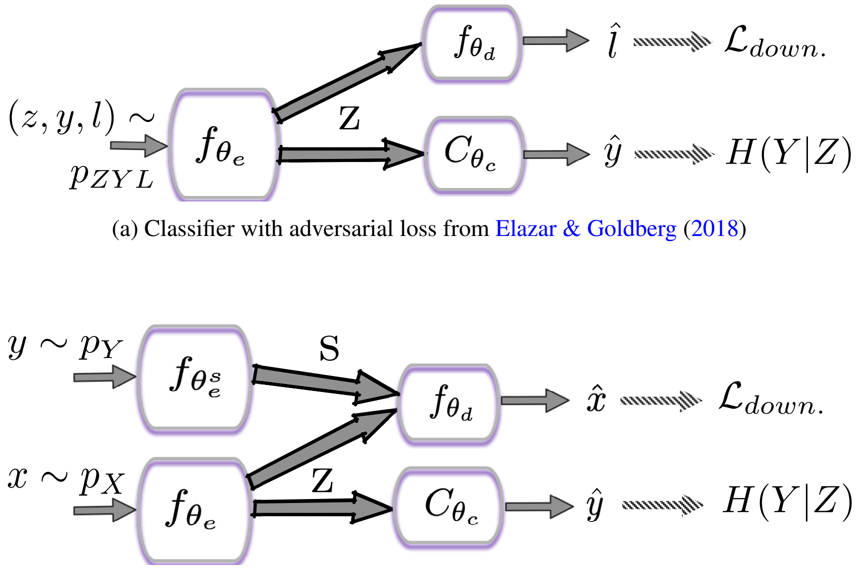
We report in Fig. 7 the schema of the baselines.

B.2 Architecture Hyerparameters

We use an encoder parameterized by a 2-layer bidirectional GRU Chung et al. (2014) and a 2-layer decoder GRU. Both GRU and our word embedding lookup tables, trained from scratch, and have a dimension of 128 (as already reported by Garcia et al. (2019), building experiments on higher

Algorithm 1 Our method for the fair classification task

```
INPUT: training dataset for the encoder $\mathcal{D}_n = \{(x_1, y_1, l_1), \ \dots, \ (x_n, y_n, l_n)\}$, batch size $m$,
    training dataset for the classifiers and decoder $\mathcal{D'}_n = \{(x'_1, y'_1, l'_1), \ \dots, \ (x'_n, y'_n, l'_n)\}$.
Initialization: parameters $(\theta_e, \theta_R, \theta_c, \theta_d)$ of the encoder $f_{\theta_e}$, classifiers $C_{\theta_R}$, $C_{\theta_c}$, $f_{\theta_d}$
Optimization:
  while $(\theta_e, \theta_R, \theta_c, \theta_d)$ not converged do
      for $i \in [1, Unroll]$ do                                        ▷ Train $C_{\theta_c}$, $C_{\theta_R}$, $f_{\theta_d}$
          Sample a batch $\mathcal{B}'$ from $\mathcal{D}'$
          Update $\theta_R$ based $\mathcal{B}'$ and using $C_{\theta_c}$
          Update $\theta_c$ with $\mathcal{B}'$
          Update $\theta_d$ with $\mathcal{B}'$
      end for
      Sample a batch $\mathcal{B}$ from $\mathcal{D}$                              ▷ Train $f_{\theta_e}$
      Update $\theta_e$ with $\mathcal{B}$ using Eq. 3 with $\theta_d$.
  end while
OUTPUT: $f_{\theta_e}$, $f_{\theta_d}$
```



(a) Classifier with adversarial loss from Elazar & Goldberg (2018)

(b) StyleEmb model from John et al. (2018)

Figure 7: Baselines methods, theses models use an adversarial loss for disentanglement. f_{θ_e} represents the input sentence encoder; $f^s_{\theta_e}$ denotes the style encoder (only used for sentence generation tasks); C_{θ_c} represents the adversarial classifier; f_{θ_d} represents the decoder that can be either a classifier (Fig. 7a or a sequence decoder (Fig. 7b). Schemes of our proposed models are given in Fig. 1

dimensions produces marginal improvement). The style embedding is set to a dimension of 8. The attribute classifier are MLP and are composed of 3 layer MLP with 128 hidden units and LeakyReLU Xu et al. (2015) activations, the dropout Srivastava et al. (2014) rate is set to 0.1. All models are optimised with AdamW Kingma & Ba (2014); Loshchilov & Hutter (2017) with a learning rate of 10^{-3} and the norm is clipped to 1.0. Our model's hyperparameters have been set by a preliminary training on each downstream task: a simple classifier for the fair classification and a vanilla seq2seq Sutskever et al. (2014); Colombo et al. (2020) for the conditional generation task. The models requested for the classification task are trained during $100k$ steps while 300k steps are used for the generation task.

C Additional Details on the experimental Setup

In this section, we provide additional details on the metric used for evaluating the different models.

C.1 Content Preservation: BLEU & Cosinus Similarity

Content preservation is an important aspect of both conditional sentence generation and style transfer. We provide here the implementation details regarding the implemented metrics.

BLEU. For computing the BLEU score we choose to use the corpus level method provided in python sacrebleu Post (2018) library `https://github.com/mjpost/sacrebleu.git`. It produces the official WMT scores while working with plain text.

Cosinus Similarity. For the cosinus similarity, we follow the definition of John et al. (2018) by taking the cosinus between source and generated sentence embedding. For computing the embedding we rely on the bag of word model and take the mean pooling of word embedding. We choose to use the pre-trained word vectors provided in `https://fasttext.cc/docs/en/pretrained-vectors.html`. They are trained on Wikipedia using fastText. These vectors in dimension 300 were obtained using the skip-gram model described in Bojanowski et al. (2017); Joulin et al. (2016b) with default parameters.

C.2 Fluency: Perplexity

To evaluate fluency we rely on the perplexity Jalalzai et al. (2020), we use GPT-2 Radford et al. (2019) fine-tuned on the training corpus. GPT-2 is pre-trained on the BookCorpus dataset Zhu et al. (2015) (around 800M words). The model has been taken from the HuggingFace Library Wolf et al. (2019). Default hyperparameters have been used for the finetuning.

C.3 Style Conservation/Transfer

For style conservation Colombo et al. (2019) (*e.g.*, polarity, gender or category) we train a fasttext Bojanowski et al. (2017); Joulin et al. (2016a;b) classifier `https://fasttext.cc/docs/en/supervised-tutorial.html`. We use the validation corpus to select the best model. Preliminary comparisons with deep classifiers (based on either convolutionnal layers or recurrent layers) show that fasttext obtains similar result while being litter and faster.

C.4 Disentanglement

For disentanglement, we follow common practice Lample et al. (2018) and implement a two layers perceptron Rosenblatt (1958). We use LeakyRelu Xu et al. (2015) as activation functions and set the dropout Srivastava et al. (2014) rate to 0.1.

D Additional Results on Sentiment

D.1 Binary Sentence Generation

D.1.1 Comparison to other work:

In Tab. 1, we report the performances of a set concurrent work as in Li et al. (2018) on 500 sentences of the test set of SYelp. It shows that our implementations reache competitive results thus validate both our implementation and our study.

D.2 Content preservation using Cosinus Similarity

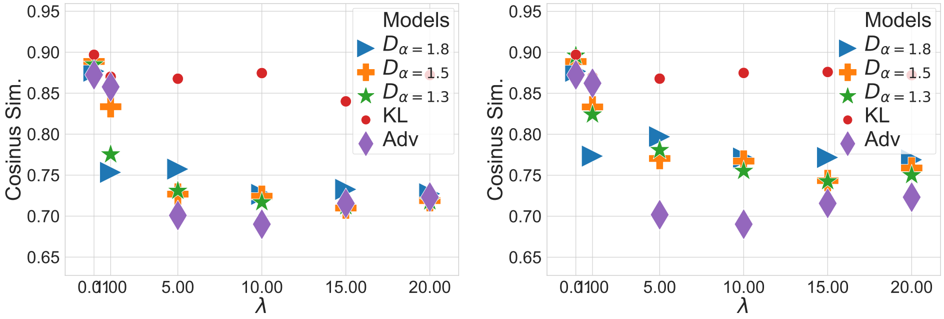
Fig. 8 measures the content preservation measured using cosinus similarity for the sentence generation task using sentiment labels. As with the BLEU score, we observe that as the learnt representation becomes more entangled (λ increases) less content is preserved. Similarly to BLEU the model using the KL bound conserves outperforms other models in terms of content preservation for $\lambda > 5$.

D.3 Example of generated sentences

Tab. 2 and Tab. 3 gathers some sentences generated by the different sentences for different values of λ.

Model	Accuracy	BLEU	PPL
MultiDecoder John et al. (2018)	54	39	5.4
Controllable Text Gen. Hu et al. (2017)	68	20	4.7
CAE Shen et al. (2017)	72	13	2.0
DeleteAndRetrieve Li et al. (2018)	76	12.5	2.1
Rule-based Li et al. (2018)	66	47	5.2
Human from Li et al. (2018)	65	31	4.2
Our: Adv, $(\lambda = 0.1/\lambda = 1)$	23/24	25/15	5.11/5.0
Our: KL, $(\lambda = 0.1/\lambda = 1)$	25/29	28/18	5.11/3.52
Our: $D_{\alpha=1.3}$, $(\lambda = 0.1/\lambda = 1)$	24/20	24.0/9.0	5.11/2.49
Our: $D_{\alpha=1.5}$, $(\lambda = 0.1/\lambda = 1)$	25/35	12.0/10.0	3.52/4.05
Our: $D_{\alpha=1.8}$, $(\lambda = 0.1/\lambda = 1)$	32/38	7.0/3.0	3.48/4.06

Table 1: Comparison with concurrent work. For this comparison we rely on the sentences provided in https://github.com/rpryzant/delete_retrieve_generate. We have reprocessed the provided sentence using a tokenizer based on SentencePiece Kudo (2018); Sennrich et al. (2016). We will release–along with our code–new generated sentences for comparison.



(a)

(b)

Figure 8: Content preservation measured by the cosinus similarity as describes in Appendix C for the Style Transfer (Fig. 8a) and the conditional sentence generation (Fig. 8b) using sentiment labels.

Style transfert. From Tab. 2, we can observe that the impact of disentanglement on a qualitative point of view. For small values of λ the models struggle to do the style transfer (see example 2 for instance). As λ increases disentanglement becomes easier, however, the content becomes more generic which is a known problem (see Li et al. (2015) for instance).

Conditional sentence generation. From qualitative example displayed in Tab. 3, we can draw similar conclusions than those for quantitative metrics previously displayed: as the disentanglement increases, the common content which is shared between input and generated sentences decreases.

Example of "degeneracy" for large values of λ. For sentences generated with the baseline model a repetition phenomenon appears for greater values of λ. For certain sentences, models ignore the style token (*i.e.*, the sentence generated with a positive sentiment is the same as the one generated with the negative sentiment). We attribute this degeneracy to the fact that the model is only trained with (x_i, y_i) sharing the same sentiment which appears to be an intrinsic limitation of the model introduced by John et al. (2018).

λ	Model	Sentence
	Input	**the food was the best food i've ever experienced.**
0.1	Adv	the food was the best i've ever had in.
	KL	the food was the best food i've ever experienced.
	$D_{\alpha=1.3}$	the food was the best food i've experienced.
	$D_{\alpha=1.5}$	the food was so good and the best i ever had.
	$D_{\alpha=1.8}$	the food is so good i will be going back.
	Input	the food was the best food i've ever experienced.
1	Adv	the food was the best i've ever eaten here.
	KL	the food was the best i've ever had.
	$D_{\alpha=1.3}$	the food was the best i've ever eaten at.
	$D_{\alpha=1.5}$	the food was amazing as well as i am extremely satisfied.
	$D_{\alpha=1.8}$	the food was very good and the service good.
	Input	the food was the best food i've ever experienced.
5	Adv	i love this place.
	KL	the food was the best i've ever eaten here.
	$D_{\alpha=1.3}$	the food is ok, but the service is terrible.
	$D_{\alpha=1.5}$	the food is always good and the service is always great.
	$D_{\alpha=1.8}$	the food was ok and very good.
	Input	the food was the best food i've ever experienced.
10	Adv	i love this place.
	KL	the food was excellent, but i love this food.
	$D_{\alpha=1.3}$	the food was best at best.
	$D_{\alpha=1.5}$	the food was well cooked with the sauce.
	$D_{\alpha=1.8}$	the food wasn't bad but it was not good.
	Input	**It's freshly made, very soft and flavorful.**
0.1	Adv	it's crispy and too nice and very flavor.
	KL	it's a huge, crispy and flavorful.
	$D_{\alpha=1.3}$	it's hard, and the flavor was flavorless.
	$D_{\alpha=1.5}$	it's very dry and not very flavorful either.
	$D_{\alpha=1.8}$	it's a good place for lunch or dinner.
	Input	it's freshly made, very soft and flavorful.
1	Adv	it's not crispy and not very flavorful flavor.
	KL	it's very fresh, and very flavorful and flavor.
	$D_{\alpha=1.3}$	it's not good, but the prices are good.
	$D_{\alpha=1.5}$	it's not very good, and the service was terrible.
	$D_{\alpha=1.8}$	it was a very disappointing experience and the food was awful.
	Input	it's freshly made, very soft and flavorful.
5	Adv	i hate this place.
	KL	it's very fresh, flavorful and flavorful.
	$D_{\alpha=1.3}$	it's not worth the money, but it was wrong.
	$D_{\alpha=1.5}$	it's not worth the price, but not worth it.
	$D_{\alpha=1.8}$	it's hard to find, and this place is horrible.
	Input	it's freshly made, very soft and flavorful.
10	Adv	i hate this place.
	KL	it's a little warm and very flavorful flavor.
	$D_{\alpha=1.3}$	it was a little overpriced and not very good.
	$D_{\alpha=1.5}$	it's a shame, and the service is horrible.
	$D_{\alpha=1.8}$	it's not worth the $ NUM.
	Input	**Only then did our waitress show up with another styrofoam cup full of water.**
0.1	Adv	then she didn't get a glass of coffee she was full full full full water.
	KL	only NUM hours of us in the water and no gratuity of a water.
	$D_{\alpha=1.3}$	waited NUM minutes at the front with us and offered to an ice glass water.
	$D_{\alpha=1.5}$	after NUM minutes of a table with a table and two entrees arrived.
	$D_{\alpha=1.8}$	after NUM minutes of a table with a table and NUM entrees arrived.
	Input	Only then did our waitress show up with another styrofoam cup full of water.
1	Adv	only NUM minutes of our waiter was able to get a refilled ice cream.

	KL	even the refund of them were brought out to refill the plate of our order.
	$D_{\alpha=1.3}$	NUM stars for the short NUM minute wait and recommend the perfect patio.
	$D_{\alpha=1.5}$	NUM minutes later, my food came out NUM minutes after our order.
	$D_{\alpha=1.8}$	i've been many years at the same time and great service.
	Input	Only then did our waitress show up with another styrofoam cup full of water.
5	Adv	great price.
	KL	she was able to get us in for a table.
	$D_{\alpha=1.3}$	they are very friendly and have a great selection of beers and drinks.
	$D_{\alpha=1.5}$	i have been here several times and it's always a good experience.
	$D_{\alpha=1.8}$	he's a great guy and a very nice person with a smile.
	Input	Only then did our waitress show up with another styrofoam cup full of water.
10	Adv	our server was very friendly and attentive.
	KL	great food, great prices, and great prices for a good price.
	$D_{\alpha=1.3}$	and i've been to this place since NUM years and love it.
	$D_{\alpha=1.5}$	only did the refill on us for about NUM mins with water tables.
	$D_{\alpha=1.8}$	i love the place.

Table 2: Sequences generated by the different models on the binary sentiment transfer task.

λ	Model	Sentence
	Input	**Definitely every flavor for every person.**
0.1	Adv	every thing have every other time.
	KL	definitely a good time to visit.
	$D_{\alpha=1.3}$	definitely worth every way every way.
	$D_{\alpha=1.5}$	definitely worth a try for all.
	$D_{\alpha=1.8}$	definitely worth a try to eat.
	Input	**Definitely every flavor for every person.**
1	Adv	definitely my wife and i love.
	KL	definitely worth every penny every time.
	$D_{\alpha=1.3}$	definitely worth the drive to earth.
	$D_{\alpha=1.5}$	definitely a recommend the whole family.
	$D_{\alpha=1.8}$	thank you for your help.
	Input	Definitely every flavor for every person.
5	Adv	definitely a good place to eat.
	KL	always a great experience.
	$D_{\alpha=1.3}$	a great place to eat.
	$D_{\alpha=1.5}$	definitely my go - to spot.
	$D_{\alpha=1.8}$	great service and great food.
	Input	Definitely every flavor for every person.
10	Adv	i love this place!
	KL	definitely get my good time there.
	$D_{\alpha=1.3}$	very good and fast service.
	$D_{\alpha=1.5}$	i would recommend this place to anyone.
	$D_{\alpha=1.8}$	definitely worth the drive.
	Input	**needless to say, i will be paying them a visit and contacting corporate.**
0.1	Adv	needless to say i will never be back with this vet... unacceptable.
	KL	needless to say i will be back and recommend this company and a complete pain.
	$D_{\alpha=1.3}$	needless to say, i will never be back to a new office and walked away.
	$D_{\alpha=1.5}$	needless to say, i will never be back to this location with my flight.
	$D_{\alpha=1.8}$	needless to say, i'm not sure what i wanted to get it.
	Input	needless to say, i will be paying them a visit and contacting corporate.
1	Adv	needless to say, i will never be back, and i am a member.
	KL	needless to say i will be back for a year and i am completely satisfied.
	$D_{\alpha=1.3}$	i wouldn't recommend this place to anyone who needs a good job.
	$D_{\alpha=1.5}$	needless to say, i will not be going back to this particular location again.
	$D_{\alpha=1.8}$	i'm not sure what i've had at this place....
	Input	needless to say, i will be paying them a visit and contacting corporate.
5		

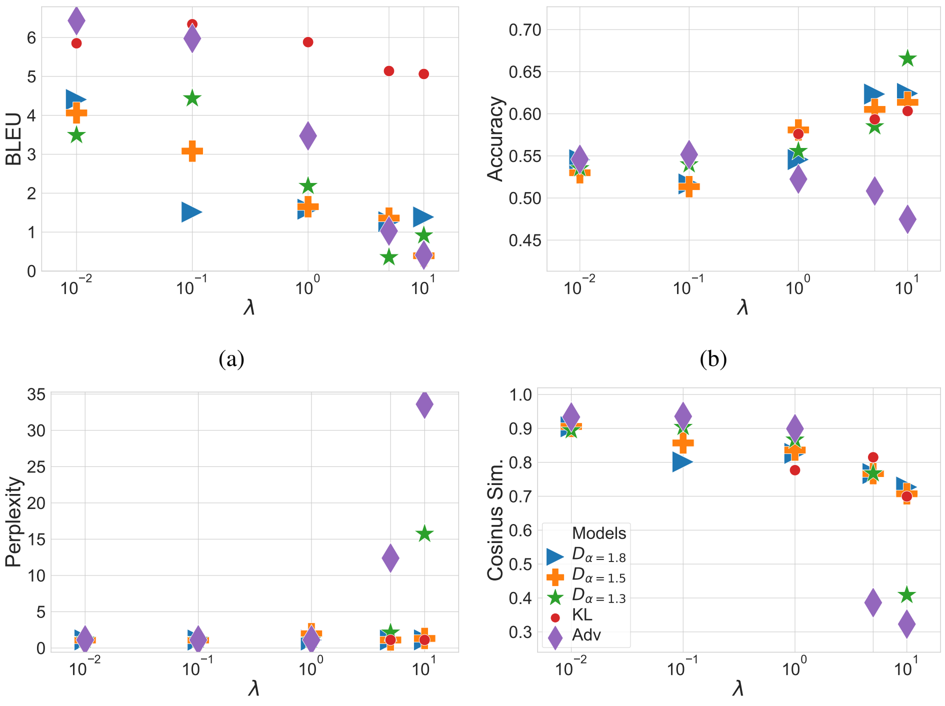
	Adv	i’m not sure what i’m going to this place.
	KL	needless to say, i will never go back, and i am completely unhappy.
	$D_{\alpha=1.3}$	they aren’t even that busy, but the food isn’t good.
	$D_{\alpha=1.5}$	if you’re looking for a good deal, you’ll find better.
	$D_{\alpha=1.8}$	needless to say, i didn’t have a bad experience.
	Input	needless to say, i will be paying them a visit and contacting corporate.
10	Adv	i’m not sure what i’ve been to.
	KL	needless to say, i will be back again, and a complete complete joke.
	$D_{\alpha=1.3}$	i’m not sure what the other reviews are to the worst.
	$D_{\alpha=1.5}$	needless to say, i will not be going back to this location.
	$D_{\alpha=1.8}$	i’ve been to this location NUM times and it’s not good.
	Input	**We had to wait for a table maybe NUM min.**
0.1	Adv	we had to wait for a table NUM mins.
	KL	we had to wait for a wait for NUM min.
	$D_{\alpha=1.3}$	we had to wait a table for NUM min.
	$D_{\alpha=1.5}$	we had a NUM minute wait for over two minutes.
	$D_{\alpha=1.8}$	we had a bad experience with a groupon for NUM.
1	Input	we had to wait for a table maybe NUM min.
	Adv	we went to wait for NUM minutes for no one.
	KL	we had a wait time for us to order NUM.
	$D_{\alpha=1.3}$	we waited for NUM minutes for a refill order.
	$D_{\alpha=1.5}$	we had a bad experience.
	$D_{\alpha=1.8}$	we had a NUM minute wait for a table.
	Input	we had to wait for a table maybe NUM min.
5	Adv	i’m not sure what i paid for.
	KL	we ordered a table for NUM minutes of our table.
	$D_{\alpha=1.3}$	we were seated immediately and we weren’t even acknowledged.
	$D_{\alpha=1.5}$	we ordered a chicken parm chicken and it was very bland.
	$D_{\alpha=1.8}$	we had a bad experience with my boyfriend’s birthday.
	Input	we had to wait for a table maybe NUM min.
10	Adv	i’m not sure what happened.
	KL	we had a table to get a table for NUM.
	$D_{\alpha=1.3}$	we ordered NUM for a lunch special and was very disappointed.
	$D_{\alpha=1.5}$	we were seated immediately and we waited.
	$D_{\alpha=1.8}$	we ordered NUM wings, NUM of NUM tacos and we waited.

Table 3: Sequences generated by the different models on the binary sentiment conditional sentence generation task.

E Binary Sentence Generation: Application to Gender Data

E.1 Quality of the Disentanglement

In Fig. 9, we report the adversary accuracy of the different methods for the values of λ. It is worth noting that gender labels are noisier than sentiment labels Lampple et al. (2018). We observe that the adversarial loss saturates at 55% where a model trained on MI bounds can achieve a better disentanglement. Additionally, the models trained with MI bounds allow better control of the desired degree of disentanglement.



(a)

(b)

(c)

(d)

Figure 10: Numerical experiments on binary style transfer using gender labels. Results include: BLEU (Fig. 10a); cosinus similarity (Fig. 10d); style transfer accuracy (Fig. 10b); sentence fluency (Fig. 10c).

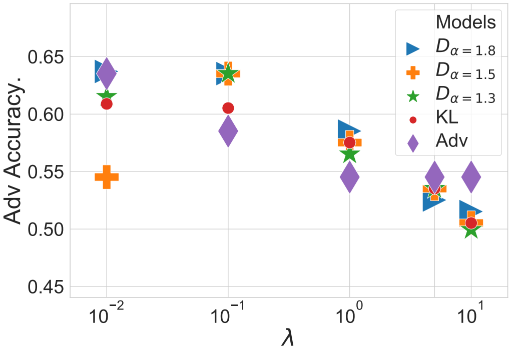
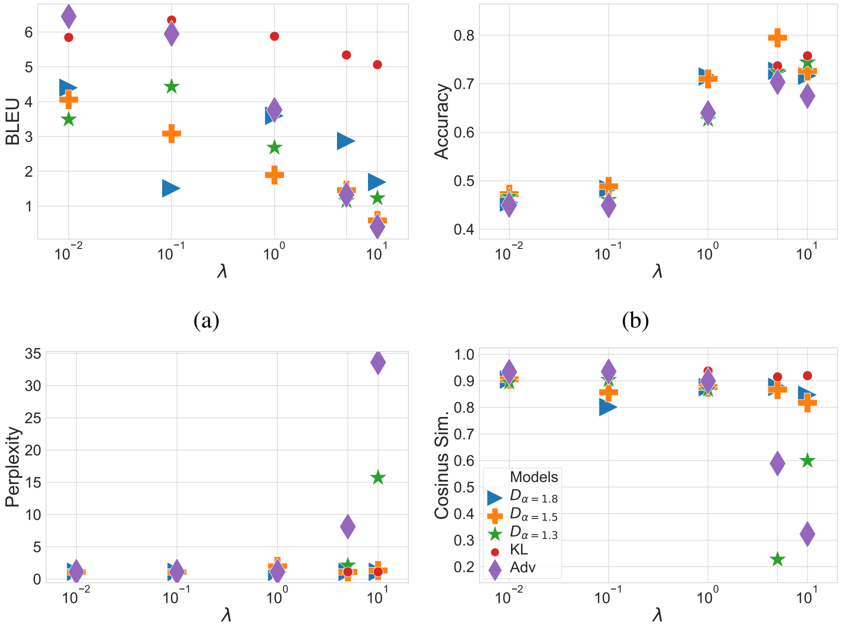


Figure 9: Disentanglement of the learnt embedding when training an off-line adversarial classifier for the sentence generation with gender data.

E.2 Quality of Generated Sentences

Results on the sentence generation tasks are reported in Fig. 10 and in Fig. 11. We observe that for $\lambda > 1$ the adversarial loss degenerates as observe in the sentiment experiments. Compared to sentiment score we observe a lower score of BLEU which can be explained by the length of the review in the FYelp dataset. On the other hand, we observe a similar trade-off between style transfer accuracy and content preservation in the non degenerated case: as style transfer accuracy increases, content preservation decreases. Overall, we remark a behaviour similar to the one we observe in sentiment experiments.



(a)

(b)

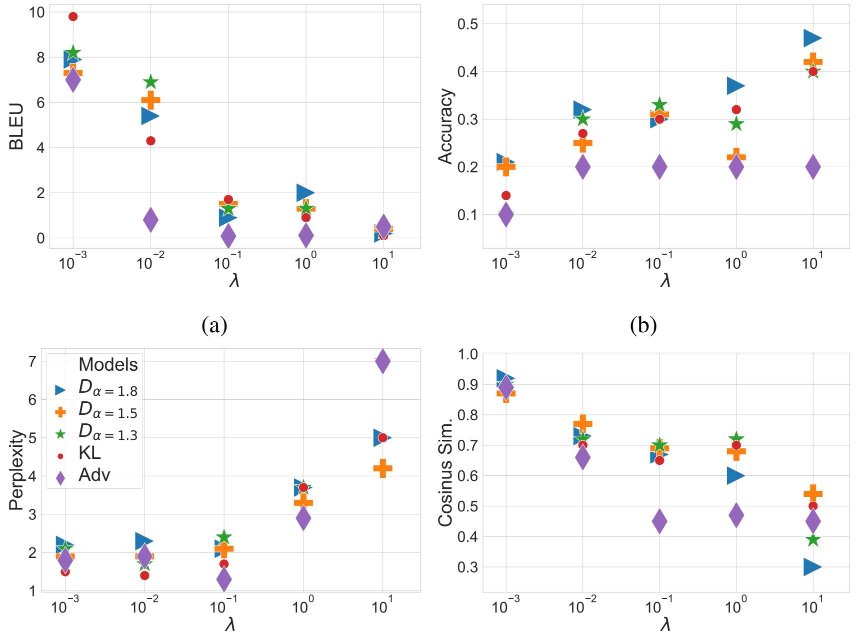
(c)

(d)

Figure 11: Numerical experiments on conditional sentence generation using gender labels. Results includes: BLEU (Fig. 11a); cosinus similarity (Fig. 11d); style transfer accuracy (Fig. 11b); sentence fluency (Fig. 11c).

F Additional Results on Multi class Sentence Generation

Results on the multi-class style transfer and on conditional sentence generation are reported in Fig. 12b and Fig. 5b. Similarly than in the binary case there exists a trade-off between content preservation and style transfer accuracy. We observe that the BLEU score in this task is in a similar range than the one in the gender task, which is expected because data come from the same dataset where only the labels changed.



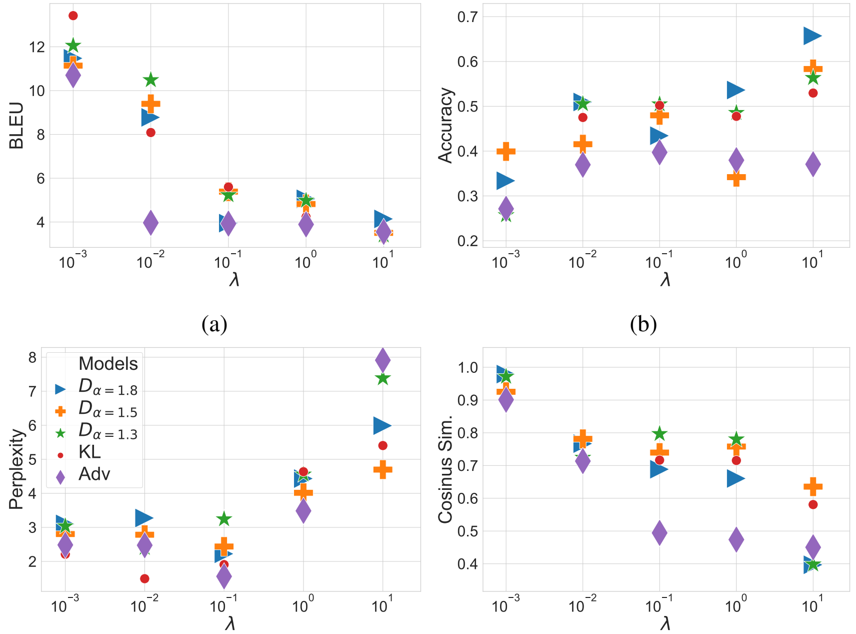
(a)

(b)

(c)

(d)

Figure 12: Numerical experiments on multiclass style transfer using categorical labels. Results include: BLEU (Fig. 12a), cosinus similarity (Fig. 12d); style transfer accuracy (Fig. 12b); sentence fluency (Fig. 12c).



(a)

(b)

(c)

(d)

Figure 13: Numerical experiments on the multi-class conditionnal sentence generation. Results include: BLEU (Fig. 13a); cosinus similarity (Fig. 13d); style transfer accuracy (Fig. 13b); sentence fluency (Fig. 13c).