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 $Q_{\hat{Y}|Z}$ be a conditional variational probability distribution on the discrete attributes satisfying $P_{ZY} \ll P_Z \cdot Q_{\hat{Y}|Z}$, 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.

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

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 $Q_{\hat{Y}|Z}$ be a conditional variational probability distribution satisfying $P_{ZY} \ll P_Z \cdot Q_{\hat{Y}|Z}$, 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

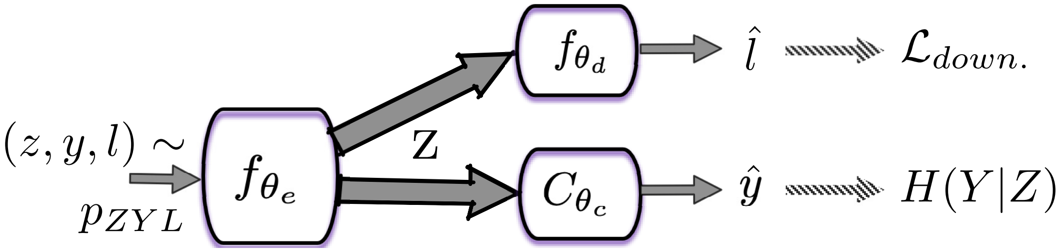
We report in Fig. 7 the schema of the baselines.

B.2 ARCHITECTURE HYERPARAMETERS

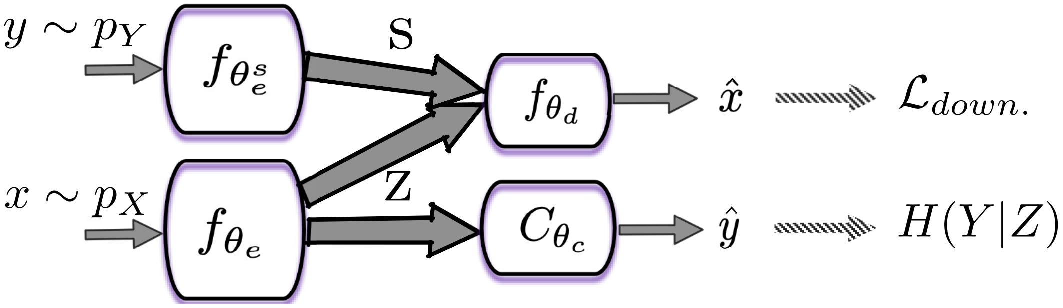
We use an encoder parameterized by a 2-layer bidirectional GRU Chung et al. (2014) and a 2-layer decoder GRU. Both GRU and our word embedding lookup tables, trained from scratch, and have a dimension of 128 (as already reported by Garcia et al. (2019), building experiments on higher

Algorithm 1 Our method for the fair classification task

```
INPUT: training dataset for the encoder $\mathcal{D}_n = \{(x_1, y_1, l_1), \ \ldots, \ (x_n, y_n, l_n)\}$, batch size $m$,
    training dataset for the classifiers and decoder $\mathcal{D}'_n = \{(x'_1, y'_1, l'_1), \ \ldots, \ (x'_n, y'_n, l'_n)\}$.
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:

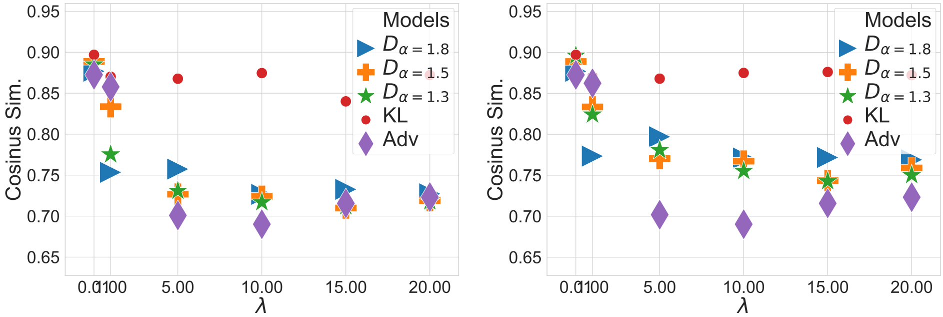
In Tab. 1, we report the performances of a set concurrent work as in Li et al. (2018) on 500 sentences of the test set of SYelp. It shows that our implementations reache competitive results thus validate both our implementation and our study.

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 $\lambda > 5$.

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.

D.3 EXAMPLE OF GENERATED SENTENCES

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

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

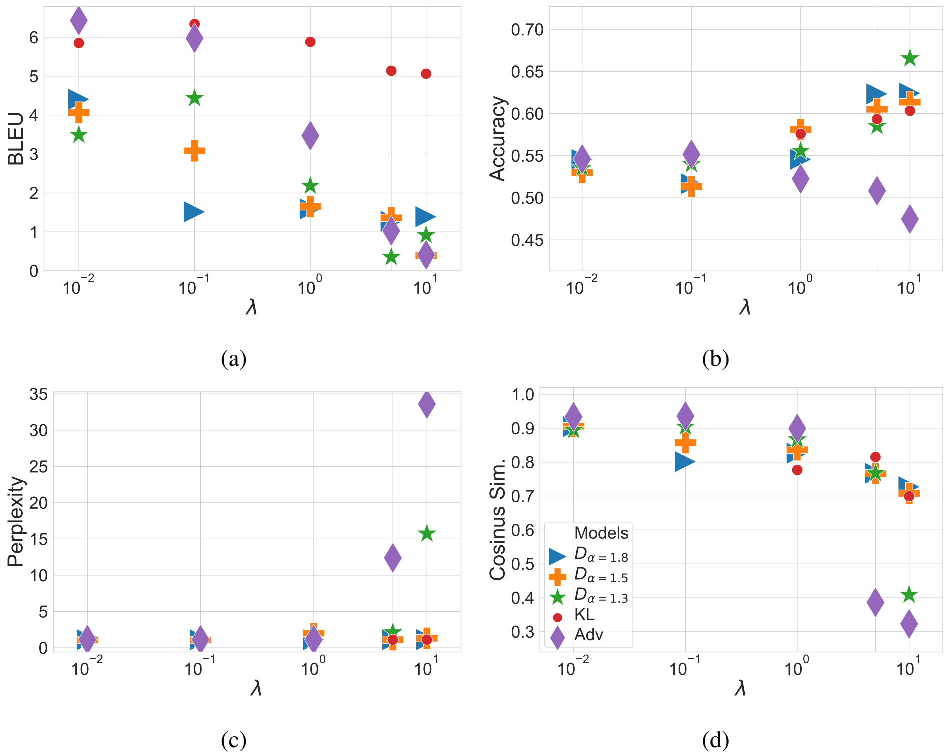
	Adv	i'm not sure what i'm going to this place.
	KL	needless to say, i will never go back, and i am completely unhappy.
	$D_{\alpha=1.3}$	they aren't even that busy, but the food isn't good.
	$D_{\alpha=1.5}$	if you're looking for a good deal, you'll find better.
	$D_{\alpha=1.8}$	needless to say, i didn't 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 Lample 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).

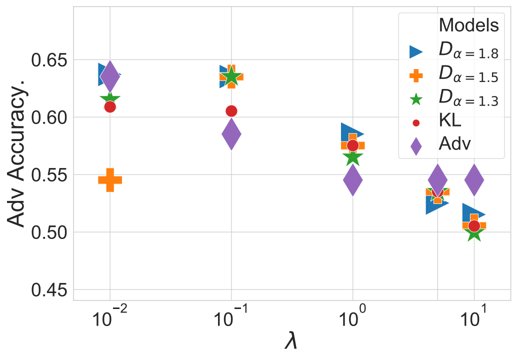
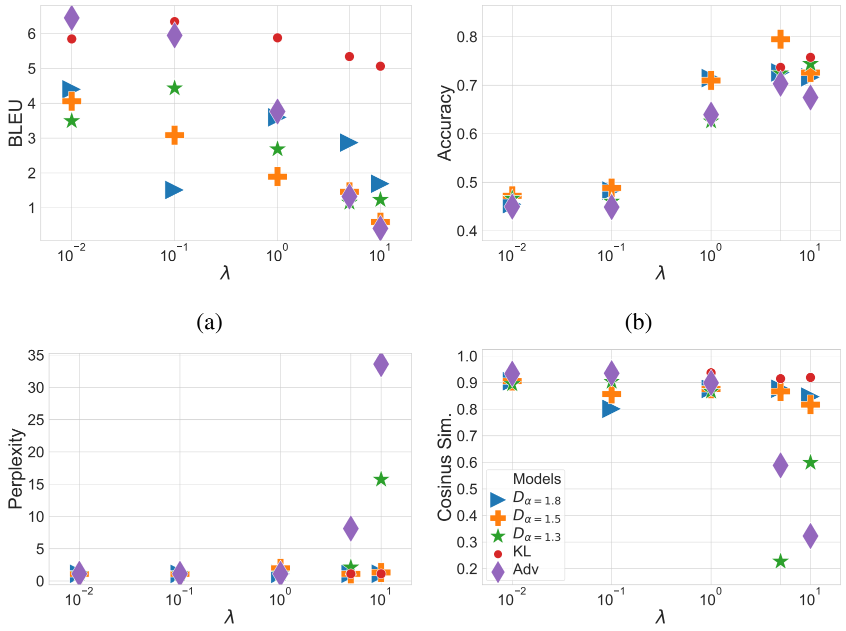


Figure 9: Disentanglement of the learnt embedding when training an off-line adversarial classifier for the sentence generation with gender data.

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.



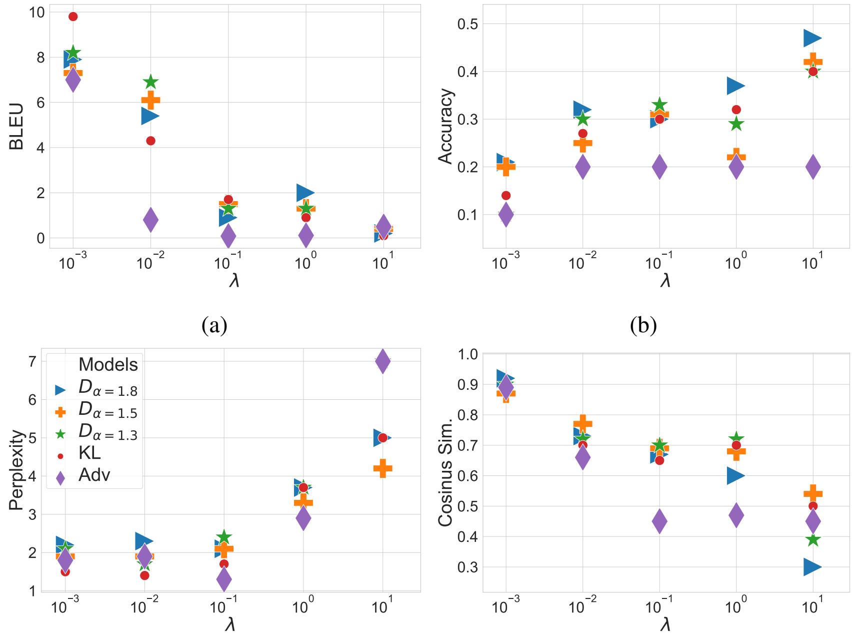
(a) (b)

(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

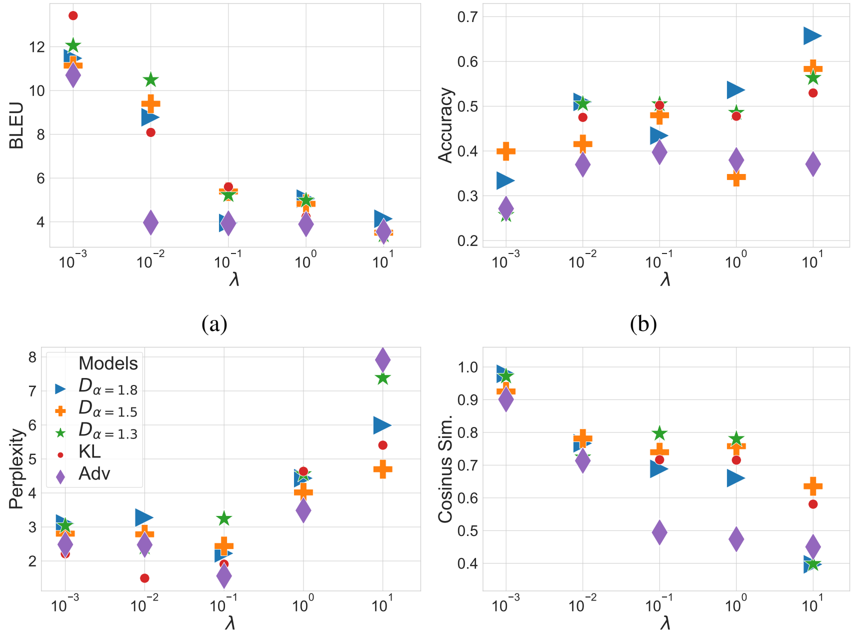
Results on the multi-class style transfer and on conditional sentence generation are reported in Fig. 12b and Fig. 5b. Similarly than in the binary case there exists a trade-off between content preservation and style transfer accuracy. We observe that the BLEU score in this task is in a similar range 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).