A. Experimental Details for Reproducibility

All the codes in this paper are implemented with PyTorch. For the implementation of β-VAE and pre-processing step in this paper, we follow the codebase of (He et al., 2019): https://github.com/jxhe/vae-lagging-encoder. The datasets used in Sec. 5 can be found in the codebase of Sudhakar et al. (2019): https://github.com/agaralabs/transformer-drg-style-transfer.

B. Details about Exploratory Experiments

B.1. Model Details for Unsupervised Sentiment Manipulation

For the β-VAE we used for the unsupervised sentiment manipulation, we use a LSTM encoding network and a LSTM decoding network. For the encoding network, the input size is 256, and the hidden size is 1,024. For the decoding network, the input size is 256, the hidden size is 1,024, and dropouts with probability 0.5 are applied on after the embedding layer and the LSTM layer in the decoding network. The dimension for the latent code is 80, and the batch size is 32. We use SGD with learning rate 1.0 to update the parameters for both the encoding and the decoding network. We train the model until the reconstruction loss stops decreasing. For aggressive training, we follow He et al. (2019) to aggressively train the encoding network. Those hyperparameters are chosen following the experiments conducted in Sec. 5 without extra tuning.

In Higgins et al. (2017), it is encouraged to set $\beta >$ to achieve better disentangling performance than vanilla VAEs. However, a large β will push the reconstruction loss higher, and the KL loss lower, leading to terrible content preservation from the source sentence. In practice, with a large β, the classification accuracy can be easily pushed to be perfect by always generating pivot sentences like "Great!" and "Terrible!". In order to achieve the best trade-off between transfer strength and content preservation, we chose β as 0.35 by inspecting the generated sentences.

B.2. Identifying the Latent Factor Indicating the Sentiment

First, we normalize the value of each latent code by subtracting the mean estimated over all the training samples. Then we use the polarity of each latent code to classify the sentiment in the validation set. The one with the highest accuracy is identified as the latent factor indicating the sentiment.

B.3. Details for β-VAE Trained on Images

For the β-VAE trained on OMNIGLOT, we use the exact same setting following the codebase of (He et al., 2019): https://github.com/jxhe/vae-lagging-encoder.

B.4. Topological Analysis: Connectedness of Mapper Graph

To help interpret the visualization in Figure 3, we give a brief description of the mapper algorithm. For a more technical introduction and details, please see Singh et al. (2007). The algorithm requires some user-specified options. The first one is a continuous function f (also called a "lens", filter or projection) that points $\boldsymbol{z}$ from a point cloud Z to $\mathbb{R}$. The range of f, $I = f(Z)$, is then divided into n overlapping open intervals $\{I_j\}_{j=1}^n$. We then find the pre-images of these intervals, $U_j = f^{-1}(I_j)$, which are open sets in the input space. Points in each U_j are then further partitioned using a clustering algorithm (e.g. DBSCAN). In the end, across all pre-images U_j's, we have a collection of clusters U_{jk}, which might or might not intersect. We represent each cluster U_{jk} as a graph node and connect two nodes if and only if the point sets intersect. Here we take the continuous function to be the sum of values in each dimension of input, and we vary n to inspect if the discovered structure persists over multiple scales or is a noise.

In the resulting graph, disconnected nodes can arise in two ways. First, if the intersecting portion of some pair of overlapping intervals does not actually contain a point mapped from the input point cloud. But this is avoided by the open cover construction in the implementation of van Veen et al. (2019). The second case is if there are actually disconnected components in the input space. Without loss of generality, assume there are two. Then by construction, some points from the two sets will be mapped to the same interval $\tilde{I}$, or shared portion of two covering intervals in the range, $\tilde{I} = I_l \cap I_k$. The pre-image of $\tilde{I}$ is the only set that could lead to a connection of the nodes, however, since it contains points that are not in the same neighborhood, clustering of this pre-image will produce two disconnected nodes, forming a disconnected graph.

B.5. Details for Topological Data Analysis

For the mapper algorithm, we use DBSCAN as the clustering algorithm. For DBSCAN, we set $\epsilon = 0.1$ and min_samples $= 3$. We sample $100,000$ points from the training set as the input. For the three cases we visualize, the latent dimensions are all 16. We choose the first 16 dimension for β-VAE trained on text and images. For CP-VAE, we use $\boldsymbol{z}^{(1)}$.

C. Details about Experiments on Text Style Transfer

C.1. Training Regimes

Across all the datasets, we use Adam with learning rate 0.001 to update the parameters for the encoding network, while SGD with learning rate 1.0 to update the parameters for the decoding network. The batch size is chosen to be 32. Dropouts with drop probability 0.5 are applied on applied on after the embedding layer and the LSTM layer in the decoding network. We train the model until the reconstruction loss stops decreasing.

C.2. Mitigating Posterior Collapse

For the structured part $\boldsymbol{z}^{(1)}$, we use β-VAE setting β as 0.2 across all the datasets. For the unstructured part $\boldsymbol{z}^{(2)}$, different strategies are employed for each dataset:

- **Yelp**: β-VAE setting β as 0.35.
- **Amazon**: β-VAE setting β as 0.35.
- **AG-News**: KL annealing, from 0.1 to 1.0 in 10 epochs.

C.3. Hyperparameter Settings

Table 6. Hyperparameter settings.

	Yelp	Amazon	AG-News
Number of variations K	3	3	10
Parameter to control the KL α	100	100	10
Input dimension for LSTM encoder	256	256	512
Hidden dimension for LSTM encoder	1024	1024	1024
Dimension for $\boldsymbol{z}^{(2)}$	64	64	96
Dimension for $\boldsymbol{z}^{(1)}$	16	16	32
Input dimension for LSTM decoder	128	128	512
Hidden dimension for LSTM decoder	1024	1024	1024

We choose $K \in \{3, 5, 10\}$, $\alpha \in \{1, 10, 100\}$, input dimension for LSTM encoder $\in \{128, 256, 512\}$, hidden dimension for LSTM encoder $\in \{512, 1024, 2048\}$, dimension for $\boldsymbol{z}^{(2)} \in \{32, 64, 96\}$, dimension for $\boldsymbol{z}^{(1)} \in \{16, 32, 48\}$, input dimension for LSTM decoder $\in \{128, 256, 512\}$ and hidden dimension for LSTM decoder $\in \{512, 1024, 2048\}$. **Amazon** follows the same setting as **Yelp** without extra tuning.

For hyperparameter tuning, one cannot rely on a single metric to tune the hyperparameters due to the inherent trade-off between transfer strength and content preservation. Instead, we search the above grids to find a setting with low reconstruction loss and high KL loss. In other words, it is likely that the setting we found is not the optimal one. Automatic approach could use multi-objective optimization to find the Pareto optimal setting. But we do not feel it is necessary here as our method is relatively insensitive to the hyperparameters except β used to control the KL loss. For β, we inspect the generated outputs on the training set to decide its value.

C.4. Decoding Strategy

For decoding, we use beam search with a beam size of 5.

D. Proof of Minimalization of Eq. 5

The problem can be formulated as an optimization problem as follows:

$$\text{maximize} \sum_{i=1}^{K} p_i^2, \quad \text{subject to} \sum_{i=1}^{K} p_i = 1.$$

By introducing a Lagrange multiplier λ, the Lagrange function is defined as

$$\mathcal{L}(p_1, p_2, \ldots, p_K, \lambda) = \sum_{i=1}^{K} p_i^2 - \lambda(\sum_{i=1}^{K} p_i - 1).$$

In order to find the optimal point, we require that

$$\frac{\partial}{\partial p_i}\left(\sum_{i=1}^{K} p_i^2 - \lambda(\sum_{i=1}^{K} p_i - 1)\right) = 2p_i - \lambda = 0, \quad i = 1, 2, \ldots, K,$$

which shows that all p_i are equal. By using the constraint $\sum_i p_i = 1$, we find $p_i = \frac{1}{K}, i = 1, 2, \ldots, K$. By plugging into the results, $\boldsymbol{\mu}^\top \boldsymbol{\mu} = \alpha \sum_i p_i^2$ reaches its minimum $\frac{\alpha}{K}$.

E. Comparisons with Baselines on Topic Modelling

Experimental setup: We use the AG news dataset for this task constructed by (Zhang et al., 2015). It contains four topic categories which are *World, Sports, Business* and *Sci/Tech*, with the title and description fields. For each category, there are $30,000$ training samples and $1,900$ test samples. In this paper, we drop the title and just use the description field. We compare our approach to two standard baselines for unsupervised topic modelling: (1) **LDA** (Blei et al., 2003), a standard implementation of LDA is used for this baseline[5]; (2) k**-means**. To show the power of our approach beyond the pre-trained sentence representations, we perform k-means clustering directly on the sentence representations. Following (Manning et al., 2010), we assign each inferred topic to one of the gold-standard topics with the optimal mapping and report the precision (*a.k.a.* purity), recall (*a.k.a.* collocation) and F_1 score. The number of topics is chosen to be 10. The results reported for the baselines and our model are the average over 10 runs.

Quantitative results: The results are shown in Table 7. We can see that our approach achieves comparable results to **LDA** while significantly outperforming k**-means** in all four categories, indicating that our approach can go beyond just clustering on pre-trained sentence representations.

Table 7. Results for topic identification.

Topic	Model	Precision	Recall	F_1
World	LDA	69.73	**75.32**	72.14
	k-means	67.64	47.63	55.90
	Ours	**80.83**	70.55	**74.59**
Sports	LDA	79.17	82.50	**80.22**
	k-means	47.66	**89.50**	62.04
	Ours	**81.14**	78.88	79.49
Business	LDA	**72.10**	66.45	**68.46**
	k-means	53.06	53.16	53.11
	Ours	64.04	**64.53**	63.97
Sci/Tech	LDA	66.55	59.77	61.60
	k-means	**81.32**	31.59	44.67
	Ours	65.20	**71.74**	**66.77**

[5]https://radimrehurek.com/gensim/

F. Text Transfer Examples

F.1. Sentiment manipulation on Yelp dataset

Table 8. Sentiment manipulation results from positive to negative

SRC	this was the best i have ever had !
B-GST	this was the worst place i have ever had !
CP-G	this was the worst pizza i have ever had !
CP-B	this was the worst i have ever had !
SRC	friendly and welcoming with a fun atmosphere and terrific food .
B-GST	the hummus is ridiculously bland and bland .
CP-G	rude and unorganized with a terrible atmosphere and coffee .
CP-B	the hummus is ridiculously greasy and tasteless .
SRC	i ordered the carne asada steak and it was cooked perfectly !
B-GST	i ordered the carne asada steak and it was just as bad !
CP-G	i ordered the carne asada steak and it was n't cooked and it was lacking .
CP-B	i ordered the carne asada burrito and it was mediocre .
SRC	the owner is a hoot and the facility is very accommodating .
B-GST	the owner is a jerk and the facility is very outdated .
CP-G	the owner is a hoot and the facility is empty and the layout is empty .
CP-B	the owner is a riot and the facility is very clean.
SRC	i will be going back and enjoying this great place !
B-GST	i wo n't be going back and this place is horrible !
CP-G	i will be going back and eat this pizza hut elsewhere .
CP-B	i will be going back and hated the worst dining experience .

Table 9. Sentiment manipulation results from negative to positive

SRC	there is definitely not enough room in that part of the venue .
B-GST	there is plenty enough seating in that part of the venue .
CP-G	there is definitely an authentic dinner in that part .
CP-B	there is definitely a nice theatre in that part .
SRC	but it probably sucks too !
B-GST	but it tastes great too !
CP-G	but it 's truly fun and insanely delicious .
CP-B	but it 's probably wonderful when u !
SRC	always rude in their tone and always have shitty customer service !
B-GST	always in tune with their tone and have great customer service .
CP-G	always great with their birthdays and always excellent music .
CP-B	always accommodating and my dog is always on family .
SRC	i was very sick the night after .
B-GST	i was very happy the night after .
CP-G	i was very pleased with the night .
CP-B	i was very happy with the night .
SRC	this is a horrible venue .
B-GST	this is a wonderful venue .
CP-G	this is a great place for celebrating friends .
CP-B	this is a great place for beginners .

F.2. Sentiment Manipulation on Amazon Dataset

Table 10. Sentiment manipulation results from positive to negative

SRC	most pizza wheels that i ve seen are much smaller .
B-GST	most pizza dough that i ve seen are much better .
CP-G	most pizza wheels that i ve seen are much more good and are much quality .
CP-B	most pizza wheels that i ve seen are much better than are much better
SRC	however , this is an example of how rosle got it right .
B-GST	however , this game is an example of how rosle loves it .
CP-G	however , this is an example of how toxic . . . sad . . . obviously .
CP-B	however , this is an example of how cheap . similar . cheap advice . cheap advice . similar .
SRC	auto shut off after num_num hours , which is a good feature .
B-GST	auto shuts off after num _ num hours , which is a shame .
CP-G	whipped mask off after num_num hours , which is slimy , which is disgusting .
CP-B	auto shut off after num_num hours, which is a stupid idea , which seems to be bad .
SRC	that said , the mic did pic up everything it could .
B-GST	that said, the game took up everything it could .
CP-G	that said, the shampoo did nt smell him well . stopped cleaning everything . ended up smelling sick
CP-B	that said, the mic did not fit everything on well , let me down it weren t cleaning
SRC	i also prefered tha blade weight and thickness of the wustof !
B-GST	i also like the blade weight and of the wustof .
CP-G	i also disliked the blade weight and thickness of the materials .
CP-B	i also slammed the blade weight and thickness of the wide .

Table 11. Sentiment manipulation results from negative to positive

SRC	the quality is declined quickly by heat exposure .
B-GST	the water is quickly drained by head exposure .
CP-G	the quality is utilitarian so grinding or sandwiches .
CP-B	the quality is priceless quickly by heat rises .
SRC	the directions were easy to follow but the quality of the easel was pathetic .
B-GST	the directions were easy to follow but the quality of the product was excellent .
CP-G	the directions were easy to follow but the quality is good for the quality and is
CP-B	the directions were easy to follow but the quality is what the quality is like the best quality of
SRC	multiplayer is just as bad, though thankfully not worse .
B-GST	quality is just as good , though thankfully not perfect .
CP-G	besides it is just good , though . those . . usually . . . usually . . .
CP-B	multiplayer is just as bad, though somebody s also so far not so far but no problem .
SRC	another energy product that simply wastes our money .
B-GST	another energy product that simply saves our money .
CP-G	another energy product that simply glides your pasta .
CP-B	another energy product that simply wastes this money .
SRC	i received the wrong color and it shreds easily .
B-GST	i received the color and it works easily .
CP-G	i low the new color and it closes easily .
CP-B	i received the wrong color and it pours easily from dishwasher and dries easily on garlic easily .

G. Text Transition Examples on AG news

Table 12. Topic transition examples.

World throughout	BAGHDAD (Reuters) - Iraq 's interim prime minister , Iyad Allawi , said on Monday that the United States had no intention of withdrawing from the country to end the violence in Iraq .
World to *Sports*	BAGHDAD (Reuters) - Iraq 's interim prime minister , Iyad Allawi , said on Monday that the United States **had no intention of withdrawing its troops from the country to the end of the year .**
World to *Business*	BAGHDAD (Reuters) - Iraq 's interim prime minister , Iyad Allawi , said on Monday that the United States **had no intention of withdrawing its troops from the country to the country .**
World to *Sci/Tech*	BAGHDAD (Reuters) - Iraq 's interim prime minister , Iyad Allawi , said on Monday that the United States **had no intention of withdrawing its uranium enrichment program to the United States .**
Sports throughout	For the first time in four years , the US men 's basketball team won the gold medal in the men 's 400-meter medley relay .
Sports to *World*	For the first time in four years , the US men 's basketball team won the gold medal at **the Athens Olympics in Athens , where the United States and the United States have agreed to a peace deal .**
Sports to *Business*	For the first time in four years , the US men 's basketball team won the gold medal at **the Athens Olympics on Wednesday , with a surge in crude oil prices .**
Sports to *Sci/Tech*	For the first time in four years , the US men 's basketball team won the gold medal in **the men 's Olympic basketball tournament in Beijing on Tuesday .**
Business throughout	NEW YORK (Reuters) - U.S. stocks opened higher on Friday , as oil prices climbed above $48 a barrel and the Federal Reserve raised interest rates by a quarter percentage point .
Business to *World*	NEW YORK (Reuters) - U.S. stocks opened higher on Friday , as oil prices climbed above $48 a barrel **and the Federal Reserve raised interest rates by a quarter percentage point .**
Business to *Sports*	NEW YORK (Reuters) - U.S. stocks opened higher on Friday , as oil prices climbed above $48 a barrel **and the Federal Reserve raised interest rates by a quarter percentage point .**
Business to *Sci/Tech*	NEW YORK (Reuters) - U.S. stocks opened higher on Friday , as oil prices climbed above $48 a barrel **and the Federal Communications Commission said it would allow the companies to use mobile phones .**
Sci/Tech throughout	SINGAPORE (Reuters) - South Korea 's Hynix Semiconductor Inc. said on Tuesday it had developed a prototype micro fuel cell recharger for a range of security vulnerabilities in India .
Sci/Tech to *World*	SINGAPORE (Reuters) - South Korea 's Hynix Semiconductor Inc. said on Tuesday it had developed a prototype micro fuel **cell aimed at ending a standoff with North Korea .**
Sci/Tech to *Sports*	SINGAPORE (Reuters) - South Korea 's Hynix Semiconductor Inc. said on Tuesday it had developed a prototype micro fuel **cell aimed at protecting the world 's biggest gold medal .**
Sci/Tech to *Business*	SINGAPORE (Reuters) - South Korea 's Hynix Semiconductor Inc. said on Tuesday it had developed a prototype micro fuel **cell aimed at protecting the world 's largest oil producer .**