On Text Style Transfer via Style-Aware Masked Language Models

Sharan Narasimhan, Pooja Shekar, Suvodip Dey, Maunendra Sankar Desarkar
Department of Computer Science and Engineering
Indian Institute of Technology Hyderabad, India

{sharan.n21,poojashekar15,suvodip15}@gmail.com
maunendra@iith.ac.in

Abstract

Text Style Transfer (TST) involves transforming a source sentence with a given style label to an output with another target style meanwhile preserving content and fluency. We look at a fill-in-the-blanks approach (also referred to as prototype editing), where the source sentence is stripped off all style-containing words and filled in with suitable words. This closely resembles a Masked Language Model (MLM) objective, with the added initial step of masking only relevant style words rather than BERT's random masking. We show this simple MLM, trained to reconstruct style-masked sentences back into their original style, can even transfer style by making this MLM "Style-Aware". This simply involves appending the source sentence with a target style special token. The Style-Aware MLM (SA-MLM) now also accounts for the direction of style transfer and enables style transfer by simply manipulating these special tokens. To learn this n-word to n-word style reconstruction task, we use a single transformer encoder block with 8 heads, 2 layers and no auto-regressive decoder, making it non-generational. We empirically show that this lightweight encoder trained on a simple reconstruction task compares with elaborately engineered state-of-the-art TST models for even complex styles like Discourse or flow of logic, i.e. Contradiction to Entailment and vice-versa. Additionally, we introduce a more accurate attention-based style-masking step and a novel "attention-surplus" method to determine the position of masks from any arbitrary attribution model in O(1) time. Finally, we show that the SA-MLM arises naturally by considering a probabilistic framework for style transfer. *

1 Introduction

Text Style Transfer can be thought of as a form of Controllable Language Generation (Hu et al., 2017) with tighter criteria. Dathathri et al. (2020) show that a classifier trained on the final embeddings of any arbitrary large language model to learn class labels of any dataset, can guide subsequent generations to resemble this dataset's style by back-propagating signals from the classifier to the LLMs activation layers. This approach, as well as other generational approaches, generally suffer from the lack of the model's ability to preserve content. This content preservation criteria, where only style influencing words/phrases must be affected, forms the critical challenge in the Style Transfer task. Learning style transfer from a parallel dataset is easy, i.e. where the output sentence for every target style is known and the model learns a straightforward mapping of input-output sentence pairs in a supervised manner. Like most work, we assume the more realistic case, where the dataset is non-parallel, and the task is unsupervised. The style transfer task is equivalent to the task of estimating the hypothetical parallel dataset from a non-parallel or "partially-observed" parallel dataset. We use this notion in later sections to elucidate the mechanisms which enable our approach.

Style is usually represented using a data-driven approach, i.e. the class labels present in an annotated dataset. One could argue that for a majority of style transfer applications, only a subset of words participate in giving the input sentence its corresponding style. As examples, we can think of this being true in the case of Sentiment, Toxicity, Formality, Politeness etc, where a few word edits can lead to a style change while the other attributes of the sentence are disaffected. As expected, this type of approach has been vastly used in the style transfer task with good success. This prototype editing approach consists a) ranking tokens according to how much they affect the underlying style b) determining which subset of tokens to mask to produce a style absent sentence, and c) transforming this style absent sentence into one that contains the

*Our code and data are available at: https://github.com/sharan21/Style-Masked-Language-Model

target style.

We show that a single self-attention encoder block trained to reconstruct style-masked sentences to the original versions using a non-parallel dataset also learns style transfer automatically. This method, resembling a masked language model, compares with state-of-the-art models albeit having a simpler training process, lesser parameters and without using pre-trained language models. We consider a novel discourse manipulation task and show that the SA-MLM outperforms another state-of-the-art model in this respect.

2 Learning Style Transfer from only Reconstruction

Many similarities appear between the editing approach for style transfer and BERT's masked language model objective. The only difference being that we reconstruct the original sentence from its style-masked version, rather than random or perplexity-based-masking sentences in typical MLMs. We postpone discussing our approach for style masking to the next section. In this section, we explore the question *"How does an MLM trained only on a reconstruction task automatically gain the ability to perform the unseen style-transfer task?"*.

An Ensemble Approach. Intuitively, we can reason that just training a model to reconstruct sentences from style-masked versions ("the food was <blank>" to the "the food was good") will not work. Since a) this does not give context to the model about styles present and b) does not give us any knob to control the output style. A simple solution would be training n different models, each to reconstruct one particular style. Style transfer can be performed by feeding a style-masked sentence to that corresponding target style model. This however a) is not computationally efficient or scale-able b) limits the learning of each model to only a fraction of the dataset, thereby over-fitting to that target domain. We perform the style transfer task using this ensemble model consisting of two generational encoder-decoder models, denoted as *Ensemble* in Tables 3, 5, 4 and 6.

A Single Style-Aware Encoder. As an alternative to training n models, we can train a single model and contextualize style by concatenating a <target_style> special token to the input style-masked sentence. Training on the reconstruction task this way (e.g. "the food was good" from "the food was <blank>. <positive>") allows it to infer the target style needed and reconstruct accordingly. This allows us to perform style transfer by simply manipulating the target style token. The model estimates the unseen portion of the hypothetical parallel dataset. Furthermore, we hypothesize that a word-by-word generational approach using an auto-regressive decoder might degrade performance in content preservation criteria. We juxtapose the performances of these approaches, denoted as *Generational* and *Encoder-only* in Tables 3, 5, 4 and 6.

Task	Positive to Negative	Contradiction to Entailment
Input	This movie is by far one of the best urban crime dramas i've seen .	a woman is sitting outside at a table using a knife to cut into a sandwich . a woman is sitting inside
Style Masked	This movie is by <mask> one of the <mask> urban crime <mask> i've seen .	a woman is sitting outside at a table using a knife to cut into a sandwich . a woman <mask> <mask> <mask>
Output	This movie is by far one of the worst urban crime garbage i 've seen .	a woman is sitting outside at a table using a knife to cut into a sandwich . a woman is a outside

Table 1: Examples of Sentiment and Discourse style transfer by the SA-MLM on the IMDb and SNLI datasets respectively.

To summarize, a simple modification (in the form of label concatenation) to the traditional MLM task enables style transfer by training solely on a reconstruction/denoising task. This forms the basis of our approach.

3 Masking Style with Attention

We explore the question, *"What constitutes a good style-masking step?"*. Intuitively, we can reason that our style-masking approach must a) produce accurate attribution scores for each token and b) use an appropriate masking policy that determines which tokens to mask using these attribution scores. The final style-masked sequences (input to the SA-MLM) must be a) completely devoid of style information and b) accurate, i.e. not done at the expense of content information.

3.1 Accurate Attribution Scores

Many prototype editing methods use Vanilla Attention (VA) as attribute scores (Wu et al., 2019b;

Zhang et al., 2018; Wu et al., 2020). It has been shown that attention is not explanation, i.e. these attribution scores do not align with human interpretability (Jain and Wallace, 2019). VA does not correlate well with other well-known attribution methods (such as Integrated Gradients Sundararajan et al. (2017)). We instead use "Explainable Attention" (EA) scores from a Diversity-LSTM classifier (Mohankumar et al., 2020; Nema et al., 2017) which have been shown to correlate better with other attribution methods as well as human judgement. We discuss more about the Diversity LSTM in section A.6 of the appendix. We also quantitatively compare the efficacy of the style-masking step between EA and VA in Table 2.

3.2 An Accurate Masking Policy

Even with having accurate attribution scores using explainable attention, effective style-masking requires careful selection of a policy which satisfies certain criteria. The primary criteria being that only tokens which significantly contribute to the style of a sentence must be masked, and other tokens must be ignored to ensure content is also preserved. Similar to the masking policy in Wu et al. (2020), it is natural to consider a "top k tokens" scheme in which the top k tokens with highest attribution are masked. However, this static approach fails for sentences which do not have exactly k style-contributing tokens, leading to either partial style masking or erroneous masking of content tokens. For the same reason, even a sentence length aware scheme such as "top 15%" masking fails. Furthermore, all such policies require sorting, leading to $O(n \log n)$ time complexity for style masking of each sentence in a batch.

Attention Surplus. Let $A = \{A_i \dots A_n\}$ denote the attention distribution of a sentence of size n. Intuitively, we can reason that all "special" tokens which might contribute more to style should have an attribution greater than the average base attribution of the sentence, given by $A^{mean} = 1/n$. Generalising this further, we refer to tokens with $A^i \geq A^{baseline}$ as tokens with "attention surplus" with respect to a sentence-length sensitive baseline attention $A^{baseline}$ given by:

$$A^{baseline} = (1 + \lambda_\epsilon) * A^{mean} \tag{1}$$

where λ_ϵ is a hyperparameter of range $0 - 1.0$. This chosen threshold $A^{baseline}$ is sensitive to the sentence length as well and subsequently ensures that the number of style-significant tokens can be dynamically determined, without need of an elaborate algorithm. As a sanity check, we observe that even in the adversarial case where all tokens might be equally important to style, A resolves into a UniformDistribution(n) and our policy correctly resorts to masking all tokens[†]. Let $Mask$ denote the token mask matrix of size n initialised with zeros.

$$\text{Mask}[A_i \geq A^{baseline}] = 1 \tag{2}$$

Using a vectorised batch-wise approach, we can style-mask an entire input batch in just O(1) complexity, compared to sorting-based approaches which take $O(\text{batch\_size} \cdot n \log n)$.

4 Related Work

Recent work focuses on various common paradigms such as disentanglement (Hu et al., 2017; Shen et al., 2017), cycle-consistency losses (Yi et al., 2020; Luo et al., 2019; Dai et al., 2019; Liu et al., 2021), induction (Narasimhan et al., 2022; Shen et al., 2020). Jin et al. (2021) and Hu et al. (2020) provide surveys detailing the current state of style transfer and lay down useful taxonomies to structure the field. In this section, we only discuss contemporary work similar to ours (prototype editing approaches) assuming the same unsupervised setting.

Li et al. (2018) present the earliest known work using the prototype editing method, in which a "delete" operation is performed on tokens based on simple count-based methods, and the retrieval of the target word is done by considering TF-IDF weighted word overlap. Malmi (2020) first train MLMs for the source and target domains and perform style transfer by first masking text spans where the models disagree (in terms of perplexity) the most, and use the target domain MLM to fill these spans. Wu et al. (2019b) introduce the Attribute-Conditional MLM, which most closely aligns to the working of the SA-MLM, also uses an attention classifier for attribution scores, a count and frequency-based method to perform masking, and a pretrained BERT model fine-tuned on the style transfer task. Lee (2020) and Wu et al. (2020) also follow roughly the same pipeline but uses

[†]Assuming $\lambda_\epsilon = 0$, whereas in practice we find $\lambda_\epsilon = 0.15, 0.5$ giving optimal masking for the Sentiment and Discourse TST, respectively. More is discussed in 7.2

a generational transformer encoder-decoder approach and also fine-tunes using signals from a pretrained classifier. Wu et al. (2019a) uses a hierarchical reinforced sequence operation method is used to iteratively revise the words of original sentences. Madaan et al. (2020) uses n-gram TF-IDF based methods to identify style tokens and modify them as either "add" or "replace" TAG tokens, which are then substituted by the decoder to perform style transfer. Similar to the SA-MLM, (Xu et al., 2018) also uses attribution-based methods from a self-attention classifier. However, they use an LSTM (Hochreiter and Schmidhuber, 1997) based approach, one to generate sentences from each domain. (Reid and Zhong, 2021) performs unsupervised synthesis to create a pseudo-parallel dataset and uses multi-span editing techniques to fill in the style using a fine-tuned pretrained language model.

5 Methodology

In this section, we describe the working of the SA-MLM during training and inference using a formal probabilistic framework.

5.1 Notation

Let S denote the set of all style labels for a annotated dataset D of the form $\{(x_0, l_0), (x_1, l_1) \dots (x_n, l_n)\}$ where x_i denotes the input sentence and $l_i \in S$ denotes the label corresponding to x_i. The set of all sentences of style s in D is denoted by $\hat{x^s} = \{x_j : \forall j \; where(x_j, s) \in D\}$. We use a special meta label m_s to represent the "style-masked" label class having s the original style. Subsequently, x^{m_s} refers to the set of all style-masked sentences with source style s. The set of all style-masked sentences from D is given by $x^m = Union(x^{m_s} : \forall s \in S)$.

5.2 A Probabilistic framework

For the sake of convenience in notation, we assume binary style labels, $S = \{0, 1\}$[‡]. We assume that a non-parallel dataset is a partially observed hypothetical parallel dataset. The SA-MLM, therefore, has to estimate the unseen half of this hypothetical parallel dataset. We follow the assumption that every output sentence with a style s is a result of:

- sampling from a latent style-masked prior, $p(x^{m_s})$,

[‡]In theory, this can be extended to any number of styles.

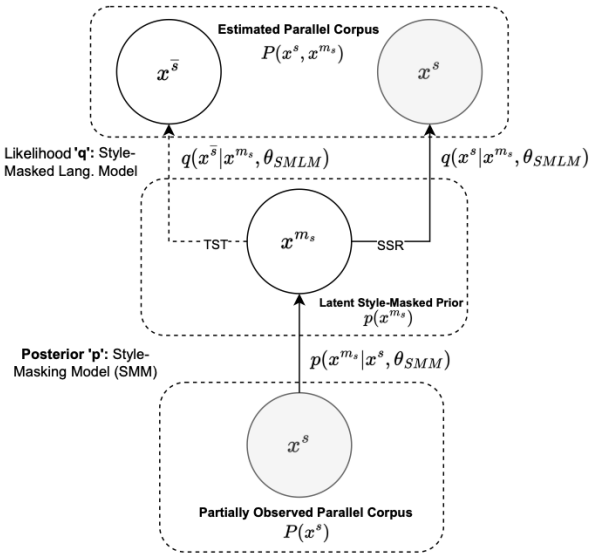

Figure 1: Probabilistic overview of our style transfer method.

- which we get from style masking input sentences, $p(x^{m_s}|x^s, \theta_{SMM})$ (posterior), and is then
- reconstructed to form the sentences with target styles using the SA-MLM, $q(x^s, x^{\bar{s}}|x^{m_s})$ (likelihood).

Style transfer is equivalent to estimating the unseen half of the hypothetical parallel $q(x^{\bar{s}}|x^{m_s})$. The overall model is summarised in Fig. 1.

5.3 SA-MLM architecture: A single self-attention block

The SA-MLM, $q(x^s, x^{\bar{s}}|x^{m_s})$ in this case, consists of a lightweight Transformer Encoder (a single self-attention block) with 2 layers, 8 heads and embeddings of size 512. To learn style transfer, training on the reconstruction task is sufficient, e.g. outting "The food was good" from a style masked input "The food was <blank>. <positive>".

5.4 Training objective

Our model, when trained only on the reconstruction task i.e. $q(x^s|x^{m_s})$, automatically learns style transfer i.e. $q(x^{\bar{s}}|x^{m_s})$. The intuition for why this is, is given in Section 2. Strictly speaking, this behaviour of automatically learning an unseen task is the result of two features we adopt, a) a single model with a common latent prior for all styles and b) the presence of target style information in the input sentences. Our model, consisting of a single

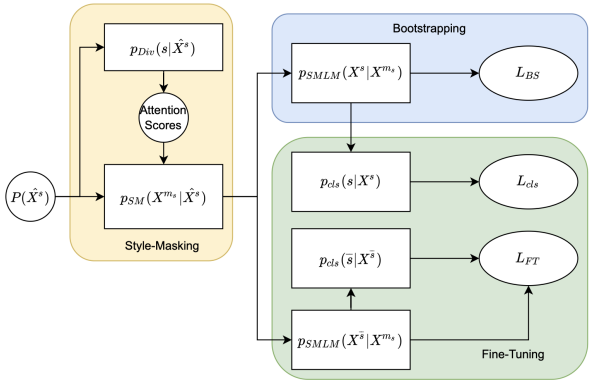

Figure 2: Overview of model architecture considering sentences of style s i.e. x^s. In reality, this is applied all styles $x^s, \forall s \in S$

self-attention encoder block, minimizes the NLL reconstruction loss.

$$L_{recon}(\theta_{encoder}) = -\log q_{encoder}(x^s|x^{m_s})$$

5.5 Fine-tuning

The presence of the special target style token during the reconstruction task forces the encoder to try and ensure that the target style is present in the output. To further enforce this behaviour, we fine-tune the encoder for one epoch using techniques similar to those found in (Liu et al., 2021). While training on the reconstruction task, we simultaneously train a classifier to predict target style labels using final layer embeddings of the reconstructed output [§].

$$L_{cls}(\theta_{cls}) = -\log p_{cls}(s|x^s)$$

For fine-tuning, this classifier then provides supervision signals to the encoder with respect to the style transfer accuracy to further enforce the outputs to align with the target style. This is formulated as min-max objective L_{FT} between the classifier cls and the $encoder$:

$$\min_{\theta_{encoder}} \max_{\theta_{cls}} -log p_{cls}(\overline{s}|x^{\overline{s}})$$

6 Datasets and Tasks

We report the split and label-wise statistics of each dataset in Table 14 of the appendix.

Sentiment Style Transfer: Following many past studies, we evaluate our model for the sentiment style transfer task using three review datasets, Yelp, Amazon and IMDb. All three datasets are annotated with two labels corresponding to positive or negative reviews and are non-parallel. [¶]

Discourse Style Transfer: Some style transfer tasks are more complex than others and have different levels of granularity (Lyu et al., 2021). To show that our seemingly simplistic approach can perform more cognitive tasks, we introduce the Discourse style transfer task by performing style transfer on the SNLI (Bowman et al., 2015) dataset. Each instance in the dataset consists of two sentences, which either contradict, entail or are neutral (no relationship). We consider the task of manipulating the discourse or "flow of logic" between two sentences, i.e. from contradiction to entailment and vice-versa. Unlike the sentiment task, which is "intra-sentence", where the style can be attributed to a select set of words, the discourse task is "inter-sentence" and requires the model to be cognizant of the context (especially for the Contradiction to Entailment Task) and detect the flow of logic.

7 Analysis of Style-Masking approach

In this section, we evaluate and justify our choice of style-masking architecture, i.e. "Explainable Attention" + "Attention-Surplus" masking policy.

7.1 Analysis of Various Attribution Methods

We consider various other attribution methods for our analysis i.e. Vanilla Gradients, Integrated Gradients (Sundararajan et al., 2017), Vanilla Attention, Attention * X (or inputs) and Explainable Attention (Mohankumar et al., 2020). We do not consider techniques such as LIME (Ribeiro et al., 2016), LRP (Bach et al., 2015), DeepLIFT (Shrikumar et al., 2017) as they are relatively more computationally expensive during inference time. For the style-masking policy, we use our "attribution-surplus" to determine which tokens to mask, as mentioned in Section 3.

In Table 2, we compute the Accuracy% and s-BLEU on the final style-masked sequences produced by each attribution method on the test split for all four datasets. We can reason that an ideal style-masking method should be able to produce sentences that completely mask out style, thereby fooling a pretrained classifier (minimizing its Accuracy%) and also preserving content information (maximizing the s-BLEU between the source

[§]We use a single feed-forward layer with input being the average of the last layer embeddings of X_i^s (excluding the meta-labels).

[¶]For Yelp, Amazon and IMDb, we used the pre-processed version specified in https://github.com/yixinL7/Direct-Style-Transfer.

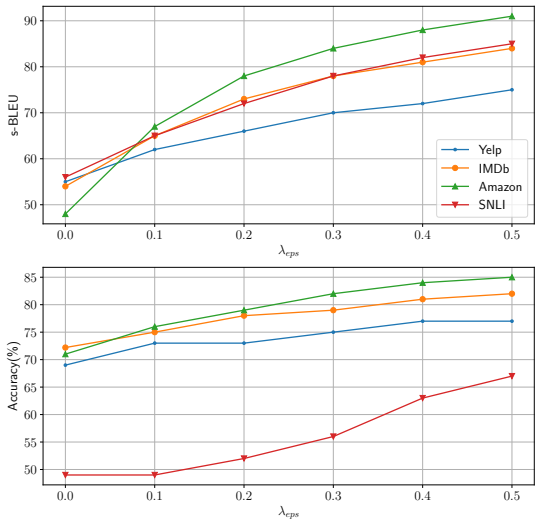

Figure 3: Effect of λ_ϵ on the resultant style-masked sentences using our "EA+AS" method. We compute s-BLEU(Top) and Accuracy%(Bottom) of the style-masked sentences on the test split of each dataset.

and style-masked sentences). We see that though Vanilla Attention is able to generally produce the lowest Accuracy%, however it does so at the expense of preserving content, reflected as lower s-BLEU compared to Explainable Attention, which, on the other hand, has the best content-preserving style masking throughout all datasets and comes as a close second in terms of Accuracy %. Other gradient-based methods do not perform favourably in any aspect.

7.2 Effect of λ_ϵ on Style-Masking

We can intuitively reason that the s-BLEU metric of the style-masked sentences serves as a rough upper bound for the s-BLEU we can potentially achieve on the output sentences after style transfer. As expected (from Eq. 1), we observe a positive correlation between λ_ϵ and both s-BLEU and Accuracy% as seen in Fig. 3. It is desirable to carefully choose λ_ϵ to be high enough to boost future s-BLEU scores on the style transfer task and also ensure that the sentences are sufficiently style masked with low Accuracy% scores. On manual checking, we observed that $\lambda_\epsilon = 0.15$ served best to accurately style-mask sentences for the Yelp, IMDb and Amazon datasets. SNLI required a higher λ_ϵ of 0.5 to ensure content information was preserved appropriately.

8 Experiments

We perform style transfer on all the datasets and analyse the results. Henceforth, we refer to the *Encoder-only* variant (described in Section 5.3) as our "SA-MLM" flagship model while comparing with baselines in Tables 3, 5, 4 and 6. The other two SA-MLM variants i.e. *Ensemble* and *Generational* described in Section 2, serve as ablation studies to support our hypothesis that a single non-generational self-attention encoder only approach is optimal for style transfer.

8.1 Automatic Evaluation Metrics

Typically, metrics for style transfer include Style Transfer Accuracy (using a pretrained classifier), BLEU for content preservation, and perplexity (using a pretrained LM) to check the fluency of outputs. Xu et al. (2020) show that this traditional set of metrics can be gamed. For fluency, we use the "Naturalness" metric (Mir et al., 2019) instead of PPL as it is shown to correlate better with human judgement. Apart from using BLEU for content preservation, we also report METEOR (Banerjee and Lavie, 2005), ROUGE-L (Lin, 2004) and CIDEr (Vedantam et al., 2015) scores in Table 12 of the appendix.

8.2 Baselines Selection

As baselines, we choose DirR (Liu et al., 2021), Stable (Lee, 2020), Transforming (Sudhakar et al., 2019), Tag (Madaan et al., 2020), CrossAligned (Shen et al., 2017), CycleRL (Xu et al., 2018), StyleEmbedding (Fu et al., 2018), D&R (Li et al., 2018) and CycleMulti (Dai et al., 2019). For the hyperparameters of each baseline, we consider the optimal parameters of the best models for each dataset reported in each respective work. Whenever available, we directly make use of the flagship style transfer outputs published as part of the original work of each reference paper to ensure that a fair comparison is done.

SOTA baselines. For all the datasets we considered, we found that the Direct (Liu et al., 2021), Lewis (Reid and Zhong, 2021) and Tag (Madaan et al., 2020) reported the strongest results out of all contemporary work in style transfer and our in experiments. We, therefore, consider these as current state-of-the-art baselines for the style transfer task to compare against.

8.3 Hyperparameter selection

The self-attention encoder block consists has 2 layers, 8 heads and embeddings of size 512. We train it for 15 epochs on the reconstruction Task and fine-tune it using signals from a pre-trained classifier

	Yelp		IMDb		Amazon		SNLI	
Attribution Model	**Acc.%**	**s-BLEU**	**Acc.%**	**s-BLEU**	**Acc.%**	**s-BLEU**	**Acc.%**	**s-BLEU**
Vanilla Attention (VA)	73.8	62.41	**69.8**	62.4	**70**	57.54	**50.76**	66
Explainable Attention (EA)	**71.3**	**64.32**	75.25	**70**	77.36	**73.21**	66.5	**85.14**
Vanilla Gradients	74.2	38.8	81.5	54.47	74.64	44.19	61.36	39
Gradients * X	97.2	37	93	50.35	84.92	40.37	70.14	39
Integrated Gradients	77.7	37.29	81.75	42.42	71	40.77	74.73	43
No Masking	100	100	100	100	100	100	100	100

Table 2: Comparison of quality of style-masking produced using various attribution models. We found that $\lambda_\epsilon = 0.0$ worked best with all gradient-based methods. For attention based methods (VA and EA), we found that $\lambda_\epsilon = 0.15, 0.5$ worked best for {Yelp, IMDb, Amazon}, SNLI respectively.

for 1 epoch. During fine-tuning, λ_{sta} is set to 1 and gradient-clipping with a threshold of 10^{-3} was set to prevent gradient explosion.

Number of Parameters. The Tag (Madaan et al., 2020) and DirR (Liu et al., 2021) models (the two best performing baselines) have 50M and 1.5B parameters respectively. The SA-MLM Encoder-only variant has 45M parameters, 30x lesser parameters than DirR's fine-tuned GPT-2 model, and roughly the same number of parameters as Tag, but outperforming it in the IMDb and SNLI datasets. We report details on training time required and infrastructure used in section A.2 of the appendix.

8.4 Quantitative metrics

We compute style transfer% (percentage of sentences with target style) using a Bi-LSTM based pretrained classifier trained on each dataset (refer section A.4 of the appendix for classifier details). r-BLEU and s-BLEU refer to the BLEU score taken between the output sentences and the human reference and ground truth sentences, respectively. For fluency, we measure the mean "Naturalness" score (the "Nat." column) as the mean classification score of a pretrained fluency discriminator [‖]. We also add a "Mean" score consisting of the average of style transfer%, s-BLEU and naturalness (normalised to 100) columns to denote a rough measure of the overall quality of each style transfer model.

8.5 Sentiment Style Transfer

Sentiment style transfer is performed on Yelp (Table 3), IMDb (Table 5) and Amazon (Table 4). We observe that for Yelp and IMDb, DirR and Encoder-only are the best-performing models according to the Mean score. In IMDb, DirR and Lewis performs better than Encoder-only in content preservation metrics but slightly lags behind in naturalness scores. In IMDb, Encoder-only achieves a significantly high style transfer% score at a reasonable s-BLEU of 60.9. In Amazon, Tag, Lewis and Encoder-only are the best performing. Overall we observe Encoder-only, DirR, Lewis and Tag as the best-performing models.

[‖]We use the pretrained naturalness classifier available in https://github.com/passeul/style-transfer-model-evaluation.

Model	**TST%**	**r-BLEU**	**s-BLEU**	**Nat.**	**Mean**
DirR	92.9	23.5	60.8	0.84	**79.27**
Stable	81.6	15.6	39.2	0.73	64.6
Transforming	84.8	18.1	44.7	0.83	70.9
Tag	87.7	16.9	47	0.83	72.57
CrossAligned	74.4	6.8	20.2	0.68	54.2
CycleRL	51.1	14.8	46.1	0.86	61.07
StyleEmbedding	8.59	16.7	67.6	0.87	54.4
D&R	88	12.6	36.8	0.89	71.27
CycleMulti	83.8	22.5	63	0.86	77.6
Lewis	93.1	-	58.5	0.84	**78.53**
Ensemble	56.5	20.5	63.2	0.85	68.23
Generational	63.4	20.3	61.3	0.83	69.23
Encoder-only	91.2	18.3	53.4	0.88	**77.6**

Table 3: Quantitative metrics for the Yelp Dataset.

8.6 Discourse Style Transfer

It is natural to suspect that prototype editing methods are only capable of working well on "course-grained" styles, i.e. where the presence of style is determined by the presence of a fixed set of words (such as "good", "delicious" in the Yelp dataset). To inspect if this is true and gauge the ability of the SA-MLM to operate on more cognitive and complex tasks, we consider "Discourse style transfer" by using Natural Language Inference (NLI) datasets.

We report statistics for the SA-MLM in Table 6. We observe that the Encoder-only model does well overall in this task and obtains a strong mean score of 85.53 (higher than the sentiment style transfer tasks). It lags behind a little mainly in the style transfer% metric, but with a strong s-BLEU and Naturalness score. We also report qualitative examples of the discourse style transfer task in Table 9 and 10 of the appendix.

Model	TST%	r-BLEU	s-BLEU	Nat.	Mean
DirR	58.2	30.1	60.6	0.91	69.93
Stable	57.2	24.9	50	0.83	63.4
Transforming	58.7	25.5	52.3	0.92	67.5
Tag	75	32.6	68.3	0.91	**78.1**
CrossAligned	73.9	1.5	2.5	0.62	46.1
StyleEmbedding	41.1	13.4	31.2	0.92	54.8
D&R	52	27.2	56.6	0.92	66.9
Lewis	74.3	-	65.6	0.93	78.53
Ensemble	52.4	31	71	0.91	**71.47**
Generational	53.4	31	69.6	0.88	70.33
Encoder-only	63.9	29.6	69.8	0.92	**75.03**

Table 4: Quantitative metrics for the Amazon dataset.

Model	TST%	s-BLEU	Nat.	Mean
DirR	80.3	67.9	0.92	**80.05**
cycleMulti	67.2	73.7	0.93	**77.95**
Ensemble	66.8	69.2	0.92	76
Generational	68.9	65.6	0.93	75.83
Encoder-only	87.9	60.9	0.92	**80.27**

Table 5: Quantitative metrics for the IMDb dataset.

8.7 Additional Content Metrics.

Past work does not tend to clarify the meaning and prioritise the presence of "content-preservation" abilities in style transfer models Lee et al. (2021). In this effort, a more thorough analysis of content preservation abilities of DirR, Tag and Encoder-only is given in section A.5 of the appendix.

8.8 Qualitative Examples

Examples of the style transfer task performed using the SA-MLM for the IMDb and SNLI dataset are given in Table 8 and 9 of the appendix.

8.9 Human Evaluations

We only consider the relatively unexplored Discourse style transfer (Entailment to Contradiction and vice versa) task for human evaluations. We were unable to reproduce the DirR and Lewis baselines to run over the SNLI dataset. Therefore, we only compare the next strongest performing models, i.e., Tag and SA-MLM. Three volunteers were given the task of voting on 200 instances (equally split for the E to C and C to E task) from the test set. A vote consists of four options, i.e., "Model 1 better", "Model 2 better" or "Both Good", "Both Bad", where the models were randomised. To determine the outcome for each instance, a majority from three separate votes was taken, one from each volunteer. In the case of no majority, the outcome is "No agreement". As seen in Table 7, the SA-MLM performs better in both tasks by a significant margin.

Model	TST%	s-BLEU	Nat.	Mean
Tag	48.3	90.2	0.98	78.83
Ensemble	52.2	88.5	0.98	79.57
Generational	58	86.7	0.98	80.9
Encoder-only	76.3	86.3	0.94	**85.53**

Table 6: Quantitative metrics for the SNLI dataset.

Direction	Tag better	SA-MLM better	Both Good	Both Bad	NA
E to C	11	**55**	8	18	8
C to E	8	**44**	2	35	11

Table 7: Human Evaluations done to compare Tag and SA-MLM on Discourse TST task on SNLI dataset. "E" and "C" denote "Entailment" and "Contradiction" respectively.

9 Conclusion

We introduce the SA-MLM, a modification of the standard MLM, which we show is capable of performing TST by using a style-masked input and performing a simple same-style reconstruction task with a lightweight Transformer Encoder block. On fine-tuning the SA-MLM for the TST%, it is on par with state-of-the-art models with orders of more parameters and sophisticated architectures in the Sentiment TST task. We show that complex styles such as flow of logic/ discourse can be manipulated even with using this simple style masking assumption. We empirically show that the SA-MLM performs well in this Discourses Manipulation task and outperforms another strong baseline in this task, also seen through human evaluations.

10 Limitations

The apparent limitation with all prototype editing models, including the SA-MLM, is that it encourages the model to only fill in necessary style words and preserve the length and structure of the original sentence. In the case of SA-MLM, the word-to-word input-output mapping while training the encoder prevents the output sentence length from changing. Though it can be argued that this even works for a relatively cognitive style like discourse, in the future, there might exist styles which explicitly require the addition/deletion of words/phrases in order to alter the style successfully. Future work will therefore focus on enabling variable-length TST outputs, similar to the (Madaan et al., 2020) approach or by incorporating a padded masked language model (Malmi, 2020).

Acknowledgements

We would like to thank all the reviewers, whose inputs and recommendations helped to substantially improve the quality of this study. We would like to thank the human annotators for their participation.

References

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE*, 10.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *IEEvaluation@ACL*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *EMNLP*.

Chandrahas, Aditya Sharma, and Partha P. Talukdar. 2018. Towards understanding the geometry of knowledge graph embeddings. In *ACL*.

Ning Dai, Jianze Liang, Xipeng Qiu, and Xuanjing Huang. 2019. Style transformer: Unpaired text style transfer without disentangled latent representation. In *ACL*.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. *ArXiv*, abs/1912.02164.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *AAAI*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9:1735–1780.

Zhiqiang Hu, Roy Ka-Wei Lee, and Charu C. Aggarwal. 2020. Text style transfer: A review and experiment evaluation. *ArXiv*, abs/2010.12742.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. 2017. Toward controlled generation of text. In *ICML*.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not explanation. In *NAACL*.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2021. Deep learning for text style transfer: A survey. *ArXiv*, abs/2011.00416.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *EACL*.

Dongkyu Lee, Zhiliang Tian, Lanqing Xue, and Nevin Lianwen Zhang. 2021. Enhancing content preservation in text style transfer using reverse attention and conditional layer normalization. In *ACL*.

Joosung Lee. 2020. Stable style transformer: Delete and generate approach with encoder-decoder for text style transfer. In *INLG*.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *NAACL*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *ACL 2004*.

Yixin Liu, Graham Neubig, and John Wieting. 2021. On learning text style transfer with direct rewards. In *NAACL*.

Fuli Luo, Peng Li, Jie Zhou, Pengcheng Yang, Baobao Chang, Zhifang Sui, and Xu Sun. 2019. A dual reinforcement learning framework for unsupervised text style transfer. *ArXiv*, abs/1905.10060.

Yiwei Lyu, Paul Pu Liang, Hai Xuan Pham, Eduard H. Hovy, Barnab'as P'oczos, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2021. Styleptb: A compositional benchmark for fine-grained controllable text style transfer. *ArXiv*, abs/2104.05196.

Aman Madaan, Amrith Rajagopal Setlur, Tanmay Parekh, Barnabás Póczos, Graham Neubig, Yiming Yang, Ruslan Salakhutdinov, Alan W. Black, and Shrimai Prabhumoye. 2020. Politeness transfer: A tag and generate approach. In *ACL*.

Eric Malmi. 2020. Unsupervised text style transfer with padded masked language models.

Remi Mir, Bjarke Felbo, Nick Obradovich, and Iyad Rahwan. 2019. Evaluating style transfer for text. *ArXiv*, abs/1904.02295.

Akash Kumar Mohankumar, Preksha Nema, Sharan Narasimhan, Mitesh M. Khapra, Balaji Vasan Srinivasan, and Balaraman Ravindran. 2020. Towards transparent and explainable attention models. In *ACL*.

Sharan Narasimhan, Suvodip Dey, and Maunendra Sankar Desarkar. 2022. Towards robust and semantically organised latent representations for unsupervised text style transfer. *ArXiv*, abs/2205.02309.

Preksha Nema, Mitesh M. Khapra, Anirban Laha, and Balaraman Ravindran. 2017. Diversity driven attention model for query-based abstractive summarization. In *ACL*.

Machel Reid and Victor Zhong. 2021. Lewis: Levenshtein editing for unsupervised text style transfer. In *FINDINGS*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Ananya B. Sai, Mithun Das Gupta, Mitesh M. Khapra, and Mukundhan Srinivasan. 2019. Re-evaluating adem: A deeper look at scoring dialogue responses. *ArXiv*, abs/1902.08832.

Tianxiao Shen, Tao Lei, Regina Barzilay, and T. Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *NIPS*.

Tianxiao Shen, Jonas Mueller, Regina Barzilay, and T. Jaakkola. 2020. Educating text autoencoders: Latent representation guidance via denoising. In *ICML*.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *ICML*.

Akhilesh Sudhakar, Bhargav Upadhyay, and Arjun Maheswaran. 2019. "transforming" delete, retrieve, generate approach for controlled text style transfer. In *EMNLP*.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. *ArXiv*, abs/1703.01365.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *EMNLP*.

Chen Henry Wu, Xuancheng Ren, Fuli Luo, and Xu Sun. 2019a. A hierarchical reinforced sequence operation method for unsupervised text style transfer. *ArXiv*, abs/1906.01833.

Chunhua Wu, Xiaolong Chen, and Xingbiao Li. 2020. Mask transformer: Unpaired text style transfer based on masked language. *Applied Sciences*, 10:6196.

Xing Wu, Tao Zhang, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019b. Mask and infill: Applying masked language model for sentiment transfer. In *IJCAI*.

Jingjing Xu, Xu Sun, Qi Zeng, Xuancheng Ren, Xiaodong Zhang, Houfeng Wang, and Wenjie Li. 2018. Unpaired sentiment-to-sentiment translation: A cycled reinforcement learning approach. In *ACL*.

Peng Xu, Jackie Chi Kit Cheung, and Yanshuai Cao. 2020. On variational learning of controllable representations for text without supervision. In *ICML*.

Xiaoyuan Yi, Zhenghao Liu, Wenhao Li, and Maosong Sun. 2020. Text style transfer via learning style instance supported latent space. In *IJCAI*.

Y. Zhang, Jingjing Xu, Pengcheng Yang, and Xu Sun. 2018. Learning sentiment memories for sentiment modification without parallel data. In *EMNLP*.

A Appendix

A.1 Ethics Statement

Any TST model can be used for illicit purposes. Therefore, it is important we keep in mind a code of ethics (e.g. `https://www.acm.org/code-of-ethics`). We will make all our code open-source and will contain all details of experimentation and implementation, training time, additional hyperparameters used in the form of log files included inside the directories of our saved models, which can also be used to replicate results.

A.2 Computational Expense and Infrastructure used

The most parameter-heavy SA-MLM model model was from the SNLI dataset. Therefore we report statistics for this model to gauge the overall computational expenses the SA-MLM demands. The model has 45 million parameters and each epoch took approximately 224 seconds to train on an Nvidia V100-SMX2 GPU and an Intel(R) Xeon(R) E5-2698 CPU. For complete details, we will make the code open source which will also contain the models we trained along with log files with all metadata about the model architecture and training.

A.3 Qualitative examples of style transfer

Qualitative examples of style transfer using the SA-MLM for the IMDb and SNLI datasets are given in 8 and 9 respectively. We also compare some qualitative examples between the SA-MLM and the Tag (Madaan et al., 2020) baselines in Table 10.

A.4 Details of pre-trained Classifier

We use a Bi-LSTM as our choice of classifier as it performs comparably to FastText (Joulin et al., 2017) and outperforms it in the SNLI dataset. A comparison of the two models is given in Table 11

A.5 Additional Content Preservation Metrics

We present more content preservation metrics in Table A.5 to compare the top three performing mod-

Direction	Negative to Positive	Positive to Negative
Input	ben affleck is back to making the same boring bad acting films .	this movie is by far one of the best urban crime dramas i 've seen .
Style Masked	ben affleck is back to making the same <mask> <mask> acting films .	this movie is by <mask> one of the <mask> urban crime <mask> i 've seen .
Output	ben affleck is back to making the same truly great acting films .	this movie is by far one of the worst urban crime garbage i 've seen .

Table 8: Example of Sentiment style transfer on the IMDb dataset.

Direction	Entailment to Contradiction	Contradiction to Entailment
Input	a guy in a red jacket is snowboarding in midair . a guy is outside in the snow	a woman is sitting outside at a table using a knife to cut into a sandwich . a woman is sitting inside
Style Masked	a guy in a red jacket is snowboarding in midair . a guy is <mask> in the <mask>	a woman is sitting outside at a table using a knife to cut into a sandwich . a woman <mask> <mask> <mask>
Output	a guy in a red jacket is snowboarding in midair . a guy is swimming in the park	a woman is sitting outside at a table using a knife to cut into a sandwich . a woman is a outside

Table 9: Example of Discourse style transfer on the SNLI dataset.

Direction	Entailment to Contradiction	Contradiction to Entailment
Input	**a black women holding a sign that says free hugs in the city . a woman is holding a sign**	**a tan dog chases a black and white soccer ball . a dog is chasing after a cat**
Output (Tag)	a black women holding a sign that says free hugs in the city . a woman is holding a sign	a tan dog chases a black and white soccer ball . a dog is chasing after a sport
Output (Encoder-only)	a black women holding a sign that says free hugs in the city . a woman is holding a book	a tan dog chases a black and white soccer ball . a dog is outside after a ball
Input	**a man is doing a task by a body of water on a farm . the man is doing something by the water**	**a dad with his child and an apple pie . a dad and his daughter with an blueberry pie**
Output (Tag)	a man is doing a nap by a body of water on a farm . the man is doing pushups by the water	a dad with his child and an apple outside . a dad and his daughter with an acousticelling outside
Output (Encoder-only)	a man is doing a task by a body of water on a farm . the man is doing nothing by the beach	a dad with his child and an apple pie . a dad and his daughter with an apple outside

Table 10: Examples of Discourse style transfer on SNLI of SA-MLM vs Tag (Madaan et al., 2020)

Dataset	Model	Acc.%
Yelp	FastText	97.6
	Bi-LSTM	97
IMDb	FastText	99.35
	Bi-LSTM	99
Amazon	FastText	92.1
	Bi-LSTM	93
SNLI	FastText	72.5
	Bi-LSTM	84

Table 11: Comparison of FastText, Bi-LSTM models for classification task on all datasets.

els i.e., SA-MLM, Tag (Madaan et al., 2020) and DirR (Liu et al., 2021).

A.6 The Diversity-LSTM and Explainable Attention

Effective style-masking requires an attribution model with a high degree of plausibility, which motivates our use of "explainable" attention scores Mohankumar et al. (2020) as choice for the style-masking step.

Why not use standard attention? Vanilla attention scores do not serve as accurate attribution scores. Attention scores over RNN hidden states for the classification task do not correlate well with other standard interpretation metrics (Jain and Wallace, 2019), such as gradient and occlusion based methods. Feeding alternative adversarial/random attention distributions lead to only a modest effects are the model's decision (Wiegreffe and Pinter, 2019). However Wiegreffe and Pinter (2019) shows that these adversarial distributions, if properly produced, do induce poorer performance showing that vanilla attention is still partially faithful to its explanation. Mohankumar et al. (2020) postulate that attention scores over hidden states (H) are not explainable due to information mixing and subsequent entanglement/coupling and mutual information among H in RNNs. To mitigate this entanglement, diversity driven learning (inspired by results in Nema et al. (2017)) is enforced among H. This promotes the attention mechanism over such diversity-enforced H to satisfy "faithfulness" and "plausibility" properties when interpreted as attribution scores, which we refer to as "Explainable attention" (EA). Mohankumar et al. (2020) empirically show that EA does not suffer any loss in performance in the downstream task. Supporting plausibility, a) EA scores correlate better with strong attribution tools such as Integrated Gradients b) On analysis over POS tags, EA attends more to tags which are contextually important w.r.t the given task and c) Correlates better to human judgement than vanilla attention.

The Diversity Driven LSTM. The Diversity LSTM consists of an LSTM-based classifier with attention (Bahdanau et al., 2015) over the H. The final context vector is fed through a feedforward layer to generate the output.

$$\tilde{\alpha}_t = \mathbf{v}^T \tanh(\mathbf{W}\mathbf{h} + \mathbf{b}) \quad \forall t \in [1, m]$$
$$\alpha_t = \text{softmax}(\tilde{\alpha}_t)$$
$$\mathbf{c}_\alpha = \sum_{t=1}^{m} \alpha_t \mathbf{h}_t$$

To enforce the H of the LSTM to be "diverse" i.e. more disentangled w.r.t each other, the conicity (Chandrahas et al. (2018), Sai et al. (2019)) metric is used as an auxillary loss and is defined as the mean of "Alignment to Mean" (ATM) for all vectors $\mathbf{v}_i \in \mathbf{V}$:

$$\text{ATM}(\mathbf{v}_i, \mathbf{V}) = \text{cosine}(\mathbf{v}_i, \frac{1}{m}\sum_{j=1}^{m} \mathbf{v}_j)$$
$$\text{conicity}(\mathbf{V}) = \frac{1}{m}\sum_{i=1}^{m} \text{ATM}(\mathbf{v}_i, \mathbf{V})$$

The attention mechanism over a Diversity LSTM's H is now encouraged to be faithful to a particular set of scores, thus promoting the model to move towards more faithful and plausible attributions. The final loss is given as:

$$L(\theta_{Div}) = -\log p_{Div}(y|P) + \lambda_c on \text{ conicity}(\mathbf{H}^P)$$

EA requires only training an additional diversity driven RNN classifier over the given dataset. After which, a single forward pass is required to obtain attribution scores. This is unlike other methods such as IG, Lime, DeepLift, Occlusion, wherein each generating each explanation requires comparatively more operations.

Diversity-LSTM Hyperparameter Selection Table A.6 gives details of the Diversity-LSTM classifier used during the style-masking step. We also note that the Diversity-LSTM's performance is comparable to a Bi-LSTM and FastText classifier (as shown in A.4).

Dataset	Model	METEOR	ROUGE-L	CIDEr	Embd. Avg. Cosine Sim.	Vector Extrema Cosine Sim.	Greedy Matching Score
Yelp	Encoder-only	0.376	0.739	4.934	0.939	0.767	0.867
	DirR	0.444	0.83	5.813	0.969	0.867	0.926
	Tag	0.362	0.707	4.326	0.934	0.765	0.867
IMDb	Encoder-only	0.414	0.8	5.657	0.96	0.755	0.891
	DirR	0.472	0.852	6.344	0.978	0.847	0.933
Amazon	Encoder-only	0.464	0.868	6.725	0.964	0.782	0.921
	DirR	0.469	0.823	6.612	0.967	0.821	0.929
	Tag	0.453	0.835	6.548	0.966	0.781	0.917
SNLI	Encoder-only	0.563	0.906	8.297	0.986	0.886	0.96
	Tag	0.606	0.944	8.619	0.992	0.921	0.972

Table 12: Content Preservation metrics for all datasets comparing top performing models

Dataset	Acc.%	$Loss_{con}$	λ_{con}
Yelp	96	0.06	10
IMDb	100	0.09	10
Amazon	89	0.03	20
SNLI	82	0.18	10

Table 13: Classification task statistics and choice of λ_{con} for each dataset.

Dataset	Style Label	Train	Dev	Test
Yelp	Positive	266K	2000	500
	Negative	177K	2000	500
IMBb	Positive	178K	2000	1000
	Negative	187K	2000	1000
Amazon	Positive	277K	985	500
	Negative	179K	1015	500
SNLI	Entailment	183K	3329	3368
	Contradiction	183K	3278	3237

Table 14: Split and label wise statistics of each dataset.