Self-Supervised Text Style Transfer with Rationale Prediction and Pretrained Transformers

Neil Sinclair[0000−0002−5869−9550] and Jan Buys[0000−0003−1994−5832]

Department of Computer Science, University of Cape Town, South Africa
sncnei001@myuct.ac.za,jbuys@cs.uct.ac.za

Abstract. Sentiment transfer involves changing the sentiment of a sentence, such as from a positive to negative sentiment, while maintaining the informational content. Given the dearth of parallel corpora in this domain, sentiment transfer and other text rewriting tasks have been posed as unsupervised learning problems. In this paper we propose a self-supervised approach to sentiment or text style transfer. First, sentiment words are identified through an interpretable text classifier based on the method of rationales. Second, a pretrained BART model is fine-tuned as a denoising autoencoder to autoregressively reconstruct sentences in which sentiment words are masked. Third, the model is used to generate a parallel corpus, filtered using a sentiment classifier, which is used to fine-tune the model further in a self-supervised manner. Human and automatic evaluations show that on the Yelp sentiment transfer dataset the performance of our self-supervised approach is close to the state-of-the-art while the BART model performs substantially better than a sequence-to-sequence baseline. On a second dataset of Amazon reviews our approach scores high on fluency but struggles more to modify sentiment while maintaining sentence content. Rationale-based sentiment word identification obtains similar performance to the saliency-based sentiment word identification baseline on Yelp but underperforms it on Amazon. Our main contribution is to demonstrate the advantages of self-supervised learning for unsupervised text rewriting.

Keywords: Text Style Transfer · Self-Supervised Learning · Transformers.

1 Introduction

Advances in large language models have enabled the generation of high-quality open-ended text [12, 20]. However, without fine-grained control, generated text has limited practical value. Despite some advances in controllable text generation, for example in writing text in a legal style [6], the transfer of text from one sentiment or style to another while maintaining the content of the source sentence [13] is still a challenging problem. Style transfer has a number of use-cases, including mitigating harmful content [22] in text, rewriting text in a more

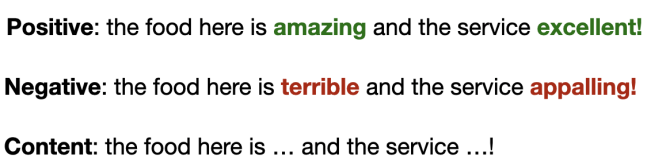

Fig. 1: An example of a sentence rewritten from positive to negative sentiment, with the sentiment words in **bold**. To rewrite the text, sentiment words are identified and removed, leaving the sentence content (bottom). The content is then rewritten in the desired sentiment.

modern style [8], or de-formalising a piece of text [21]. However, one of the key challenges that differentiates sentiment and style transfer from other text transduction tasks is the lack of parallel corpora.

In this paper we extend previous unsupervised approaches to text sentiment transfer by utilizing self-supervised learning for rewriting sentences. Our approach extends one of the main paradigms to text sentiment and style transfer [13, 29, 23] that identifies and deletes sentiment-specific words and learns sentiment transfer through sentence reconstruction conditioned on the target sentiment (Figure 1). Our work builds on previous approaches in three ways: First, we utilise the method of rationales [11, 1], a neural network-based approach from the interpretability literature, to identify and mask sentiment words. This replaces the previous heuristic n-gram saliency approach [13, 29]. Second, we fine-tune a pretrained BART [12] model to reconstruct masked sentences, which enables generating sentences with a different sentiment autoregressively. BART, pretrained with a denoising autoencoder (DAE) objective, is a natural fit for sentence reconstruction training. Third, we use self-supervised training to further improve the model's performance utilising its own generations: The fine-tuned model is used to generate a high-precision parallel corpus of sentences in opposite sentiments, on which BART is fine-tuned further to improve its style transfer accuracy.

We evaluate our approach on Yelp and Amazon review datasets [4]. We compare rationale-based sentiment word identification to the saliency-based approach, and BART to a sequence-to-sequence (Seq2Seq) [25] model with Long short-term memory (LSTM) [5]. Results using both automatic and human evaluations show that rationale-based sentiment word identification performs on par with the saliency-based approach on the Yelp dataset, but underperforms it on the Amazon dataset when used with the BART model. On the Amazon dataset BART obtained higher BLEU score but reduced sentiment transfer accuracy compared to the Seq2Seq model. Self-supervised training improves sentiment transfer accuracy on both datasets. Rationale-based sentiment word identification obtains similar performance to the saliency-based sentiment word identification baseline on the Yelp dataset while offering fine-grained control over the trade-off between content preservation and sentiment transfer accuracy. However it underperforms on Amazon due to its structure which makes it harder to identify style words. BART outputs were rated higher by human evaluators than the

sequence-to-sequence models' on both datasets, in particular due to better fluency. The performance of our approach is close to that of state-of-the-art models [29, 27] on Yelp, but lower on Amazon.

2 Background

The goal of text style transfer is to change some attribute of a sentence, such as its style or sentiment, while maintaining its content [13]. Due to the lack of parallel corpora across different styles in most domains [23], text style transfer is usually approached as an unsupervised learning problem, learning from non-parallel examples of text in different styles.

There are two main paradigms of approaches to sentiment and style transfer. In the first paradigm, sentences are seen as a combination of content and style elements. The style words are removed and the model is trained to reconstruct the full sentences from the content elements plus a token representing the style of the sentence. The model is effectively a semi-supervised denoising autoencoder (DAE), where the style words are removed based on some pre-defined criteria. The Delete Retrieve Generate (DRG) [13] model implements this paradigm by using a heuristic n-gram saliency-based approach to identify style-specific words. A sequence-to-sequence LSTM is trained with a DAE objective to reconstruct the original sentence. Additionally, the "retrieve" step retrieves relevant words or sentences in the target style, and conditions on this when generating sentences in the target style. We use the delete-only version of DRG as a baseline in this paper, as our BART-based model is trained similarly without a retrieval step. Figure 1 gives an example of sentiment transfer by deleting sentiment words and then rewriting the sentence in the target sentiment.

Subsequent work extends DRG to use a Transformer [26] instead of an LSTM while utilising the attention weights of the model to identify which words to remove [23], and uses a pretrained BERT [2] model that learns to fill in masked out style words [29]. This formulation simplifies learning but it limits the application of style transfer to be narrowly defined as deleting and replacing words.

The other paradigm to text style transfer involves encoding the input sentence in a latent representation, manipulating the encoded representation, and then decoding the sentence in another style. Wang et al. [27] utilise a Transformer to learn a latent representation of a sentence comprising both style and content, and then use a pretrained classifier to edit the entangled latent representation. Logeswaran et al. [14] and Lample et al. [10] approach style transfer as a combination of autoencoder (reconstruction) and back-translation objectives, while He et al. [3] apply variational inference to generalise these approaches.

3 Sentiment Word Identification

We review a widely-used approach to sentiment word identification based on n-gram saliency, and introduce our approach to apply an interpretable neural classifier to identify and mask or remove sentiment words.

Let $D = (x^{(1)}, c^{(1)}, ..., (x^{(m)}, c^{(m)})$ be the training set of sentences $x^{(j)}$ each annotated with sentiment marker $c^{(j)}$, where in our application the sentiments are restricted to $c^{(j)} \in \mathcal{C} = \{\text{positive}, \text{negative}\}$, and D_c denotes the subset of sentences in sentiment c.

3.1 Saliency Noising

DRG [13] uses a heuristic approach to identifying sentiment-specific words. This approach is conceptually similar to term frequency inverse document frequency (TF-IDF) where words that have higher discriminative power are weighted higher. The *salience* of word or n-gram w with respect to sentiment c is calculated as

$$s(w, c) = \frac{\text{count}(w, D_c) + \lambda}{(\sum_{c' \in \mathcal{C}, c' \neq c} \text{count}(w, D_{c'})) + \lambda}, \tag{1}$$

where λ is a smoothing parameter. The method identifies w as a sentiment marker for sentiment c if $s(w, c) > \gamma$, where γ is the saliency threshold.

3.2 Rationales Noising

Motivation Interpretable neural network-based text classifiers identify which words are most important for a classifier's classification decision for a given input. Gradient-based methods such as Integrated Gradients [24] sum the gradients across input features in order to understand which parts of the input are most sensitive to changes in the output. In this manner it uncovers the contribution of each feature to the output, enabling identifying the features or rules the model uses to make decisions.

Another approach for interpretable neural networks is the method of rationales [11]. This method aims to understand how a neural network-based classification model reaches its prediction by identifying the words or phrases that are essential for the model to achieve the correct sentence classification. The model learns a binary mask over the input sentence that learns to mask words which do not contribute to the rationale. Bastings et al. [1] extends this work using the Hard Kumaraswamy distribution to enable the differentiable binary masking of tokens. In contrast to Integrated Gradients, which provides a real number representing the contribution of each feature to a model's output, the method of rationales identifies a discrete set of tokens most impacting the classification.

In the context of sentences written in a particular sentiment, the words that are most important in classifying its sentiment will most likely be the sentiment words of a sentence. We therefore propose a novel approach to identifying sentiment words based on interpretable classifiers.

Method Given an input sentence x, the method of rationales [11, 1] defines a latent binary variable Z_i corresponding to each sentence token x_i, indicating whether x_i is included in the rationale for the classification decision. For each

sentence, the tokens in x that do not form part of the rationale are masked out, while the unmasked words are used as input to a trainable classifier that predicts the sentiment marker c. Masking can be formulated as taking the element-wise Hadamard product of the latent variable and the sentence tokens, i.e., $z \odot x$.

The rationales (masks) Z_i can each be seen to be derived from a Bernoulli distribution, and the sentence classification variable C from a categorical distribution:

$$Z_i|x \sim \text{Bern}(g_i(x;\phi)), \tag{2}$$

$$C|x,z \sim \text{Cat}(f(x \odot z;\theta)). \tag{3}$$

Functions $g_i(x;\phi)$ and $f(x \odot z;\theta)$ are neural networks based on the LSTM [5] architecture and are parameterised by ϕ and θ respectively.

To enable differentiable training with respect to a discrete set of masks, a modified version of the Kumaraswamy distribution [9] called the Hard Kumaraswamy distribution is employed [1]. The reparameterization trick [7] is utilised during training. The per-example loss function is defined as

$$-\mathbb{E}_{P(z|x,\phi)}[\log P(c|x,z,\theta)] + \lambda \sum_{i=1}^{n} z_i, \tag{4}$$

where λ is a hyper-parameter. The first part of the loss is the lower bound on data log-likelihood $\log P(c|x)$, the likelihood of the sentiment label given the sentence x. The second part is a sparsity loss to prevent the model from choosing the entire sequence as the rationale. We do not include an additional loss used by [1] that encourages choosing longer contiguous stretches of words.

4 Self-Supervised Text Sentiment Transfer

The training pipeline for our approach comprises five steps: training a classifier, training and applying the rationales sentence noiser, training the BART DAE, generating a high-precision parallel corpus, and finally self-supervised training. The training process is visualized in Figure 2.

Classifier Training A BART encoder is fine-tuned to classify the sentiment of sentences in the training corpus (step 1 in Figure 2). This classifier is used for self-supervised training.

Sentence Noising The rationales model [1] is trained as a (second) sentiment classifier. The extracted rationales are represented as a vector of binary masks for each sentence. Each token that is part of the rationale is replaced with a <mask> token or removed in the case of the sequence-to-sequence model. This process is shown in step 2 of Figure 2. We refer to the masking or removal of sentiment-specific words from the sentences used as input to the sentiment transfer models as *noising*, as both the BART and seq2seq models are trained as DAEs. We train rationale extraction models with different noising levels, and as sentiment word identification baseline we also use saliency-based noising.

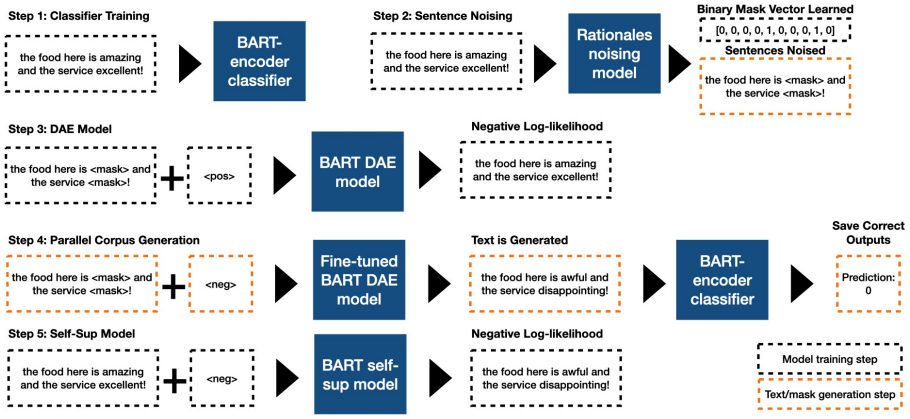

Fig. 2: Overview of classifier pretraining, sentence noising, and DAE training in the BART training pipeline.

DAE Model Training BART [12] is a pretrained encoder-decoder Transformer that can be seen as a generalisation of BERT [2] as the encoder and GPT [19] as the decoder. BART is trained with a DAE objective in which encoder input tokens are randomly masked (and permuted) and the input is reconstructed autoregressively by the decoder.

In our approach the BART encoder-decoder is fine-tuned as a DAE to reconstruct the sentences in the training corpus in which the sentiment words have been masked with a <mask> token, using either the rationale-based or saliency-based method. No word permutation is done. The sentiment token (<pos> or <neg>) is appended as the first input token in the decoder. This is shown as step 3 in Figure 2.

The model is trained to optimise the negative log-likelihood of each of the sentences that it aims to reconstruct:

$$L_{\text{DAE}} = -\sum_{j=1}^{m} \log p(x^{(j)}|\text{Mask}(x^{(j)})), \tag{5}$$

where $\text{Mask}(x)$ represents the sentence with the sentiment words masked out.

Early stopping is performed based on the accuracy of validation sentences translated to the opposite sentiment, as measured by the classifier trained in step 1. Every fixed number of batches 1 000 sentences from the development set are sampled and masked and fed to the partially fine-tuned BART model to translate by seeding the decoder with the sentiment token opposite of the original sentence. Sentiment transfer accuracy is measured by feeding the generated output sentences to the classifier trained in step 1.

Parallel Corpus Generation We apply the fine-tuned BART model to generate a parallel training set of sentences in opposite sentiments by translating all

the (masked) sentences from the training set. This is shown as step 4 in Figure 2. Sentences are generated auto-regressively using greedy decoding, where the next word with the highest probability of occurring is chosen deterministically. Minor formatting corrections such as removing double spaces and some excess punctuation is also necessary. The generated sentences are filtered so that only sentences classified as having been accurately transferred into the target sentiment (using the classifier from step 1) are kept.

Self-Supervised Training Finally, the BART model is fine-tuned further to translate sentences from one sentiment to another with self-supervised training (step 5 in Figure 2). Here, we use the term self-supervised as the parallel training set was generated and filtered entirely by the model. In this training step original (unmasked) sentences are used as inputs, as we found that the model obtains higher sentiment transfer accuracy in such a setting. The same setup is followed at test time.

The self-supervised training loss is defined as

$$L_{\text{SS}} = -\sum_{(x,y,c,c')\in D'} \log P(y|x,c'), \tag{6}$$

where D' represents the new dataset of generated parallel sentences with y as the generated sentence translated from sentiment c to sentiment c'. The model is trained until the early stopping criterion is met where translation accuracy does not improve over three validation checks of 250 batches each.

5 Experimental Setup

5.1 Data

We use two common sentiment transfer datasets, which are truncated versions of Yelp and Amazon reviews that have been categorised as either positive or negative [4]. The training sets have 440k and 555k training examples, and an average sentence length of 7.9 and 13.8 tokens, respectively. The test sets include human-annotated reference translations which are used to calculate the automatic evaluation Bilingual Evaluation Understudy (BLEU) scores [15]. These scores measure the reliability of the content of the generated sentences in the new sentiment with respect to a gold standard set of sentences.

5.2 Sentiment Word Identification

We use the original implementation of the rationales model [1] to learn sentiment token identification.[1] To test the impact of different levels of token masking or removal on sentiment transfer accuracy and content consistency of the generated sentences, five levels of rationales-based noising are used, namely 15%, 20%, 30%,

[1] https://github.com/bastings/interpretable_predictions

40% and 50%. There is some variance in the actual level of noising achieved per sentence, and also some discrepancy between the mean level of noising constraint fed to the model and the mean level of noising achieved.

As a sentiment word identification baseline we use a PyTorch implementation of the saliency-based method [18].[2] Saliency noising achieves a mean level of token masking or removal of 32.4% on Yelp and 31.1% on Amazon, so it is closest to the rationales 30% noising scheme. However its inter-sentence variance is higher than that of rationales noising.

The rationales model [1] is trained for 20 epochs with an Adam optimizer with learning rate set to 2e-4, and a batch size of 128. Pretrained GLoVE embeddings [16] are used with an embedding size of 300. For saliency noising the smoothing parameter, λ is set to 1; spans of up to four words are considered as attribute markers; and γ, the threshold for the saliency score, is set to 15 and 5.5 for Yelp and Amazon respectively. These hyperparameters follow [13], in which they were chosen through tuning on the development set.

5.3 Sentiment Transfer Models

We use the HuggingFace Transformers [28] pretrained implementation of BART-Base with a language modelling head. Due to a degree of stochasticity in the results each version of the model was trained three times and the final accuracy and BLEU results of the three runs averaged on the test set for reporting. During validation and testing, all sentences are generated using greedy decoding. This lead to better performance than sampling-based decoding, where each word is sampled based on the next word probability distribution.

BART-Base consists of six encoder and decoder blocks in the encoder and decoder with 139 million parameters in the model. The dimensionality of the model is $d_{model} = 768$. The weights are optimized using the Adam optimizer with a learning rate of 2e-5 which was empirically found to achieve the best results. We use early stopping if the translation accuracy fails to improve for 3 consecutive validations, which are conducted every 250 batches. This is to prevent model over-fitting and is used due to the model appearing to learn very quickly from the data. A batch size of 64 is used for training due to memory constraints. All the parameters are fine-tuned except the positional embeddings.

As a non-pretrained baseline we use the delete-only Seq2Seq model of [13], utilising the PyTorch implementation of [18]. This model is also trained with various levels of token masking using the rationales model, in addition to using the saliency method. This model uses word-based tokenization with a vocabulary size of 16,000, in contrast to the word-piece tokenization used in BART. The model is trained utilising an Adam optimizer with learning rate of 2e-4. The model is trained for 70 epochs for both datasets. The embedding dimension is kept at the default of 128 and hidden dimension of the LSTM encoder and decoder is kept as 512 as per the original paper.

[2] https://github.com/rpryzant/delete_retrieve_generate

Table 1: Classification accuracy and BLEU score of BART and Seq2Seq models for different levels of token masking or removal for the Yelp and Amazon datasets.

	Yelp				**Amazon**			
	BART		Seq2Seq		BART		Seq2Seq	
Noising	Accuracy	BLEU	Accuracy	BLEU	Accuracy	BLEU	Accuracy	BLEU
15%	70.3	29.4	49.0	24.3	28.5	37.4	39.2	19.5
20%	74.5	27.7	69.2	16.1	31.4	33.5	40.0	29.6
30%	78.6	26.3	72.0	17.8	35.2	27.2	42.4	22.3
40%	83.5	21.1	81.2	13.6	48.1	23.1	51.7	16.4
50%	93.7	13.6	92.0	8.74	44.7	18.1	58.8	11.0
Saliency	74.2	26.2	82.2	15.1	53.6	37.0	47.4	21.3
Wu et al. (2019)	97.3	14.4	-	-	84.5	28.5	-	-
Wang et al. (2019)	95.4	22.6	-	-	85.3	34.1	-	-

6 Results

6.1 Automatic Evaluation

We evaluate our approach based on the accuracy of rewriting sentences from one sentiment to another as assessed by the pretrained classifier, as well as by the BLEU score [15] comparing generated sentences with a set of reference human translated sentences.[3] BLEU represents the outputs' content preservation, and to a lesser extend their fluency. For both models sentiment word identification is performed with rationale-based noising with different noising levels as well as with the saliency-based sentiment word identification baseline.

Table 1 gives the results of both our BART-based self-supervised model and our replication of the delete-only Seq2Seq DAE model of [13]. The table reports BART test set results with the full training pipeline including self-supervised learning. Additionally we include the results of Wu et al. [29] and Wang et al. [27], which are state-of-the-art on the Yelp and Amazon datasets, respectively. As an ablation experiment we found that BART with only DAE training obtains style transfer accuracies that are between 4.3% and 20.4% lower than including self-supervised training (across different levels of rationales noising on both datasets). This highlights how effective the models are at learning the sentiment of a sentence conditioned on a sentiment token. This is especially evident as sentence noising or token masking increases. However, it also highlights the limitations of training without the paired synthetically generated self-supervised training set. These gains in accuracy are likely due to the model learning an explicit mapping from one sentiment to another when using the synthetically generated training set. This is in contrast to learning to simply replace masked words in a sentence conditioned on a sentiment token.

[3] We use SACREBLEU [17] with default settings to calculate the BLEU score.

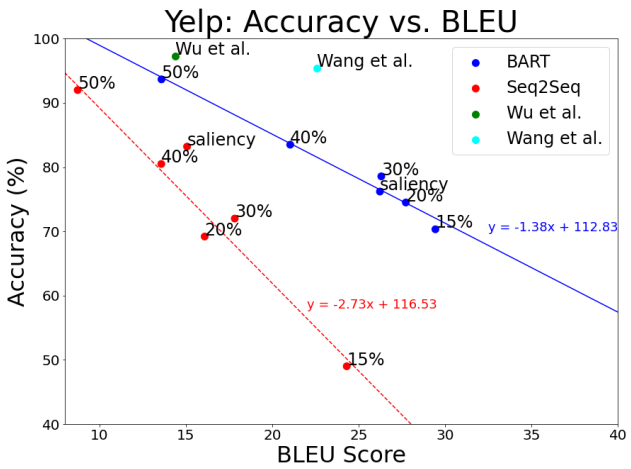

Fig. 3: Test set sentiment transfer vs. BLEU score for BART (solid blue line) and Seq2Seq (striped red line) on the Yelp dataset.

The results on the Yelp dataset shows that our BART-based model achieves higher accuracy and higher BLEU scores than the Seq2Seq model for the same level of rationale-based noising. Saliency-based noising leads to a higher BLEU score but lower classification accuracy on BART compared to the Seq2Seq baseline. On the Amazon dataset BART achieves higher BLEU but lower accuracy than Seq2Seq for all the levels of rationale-based noising. Saliency-based noising leads to the highest overall trade-off between accuracy and BLEU score, as can be seen in Figure 4. Seq2Seq with 50% rationale-based noising, however, obtains the highest overall classification accuracy.

The trade-off between sentiment transfer accuracy and content preservation as measured by BLEU is visualised in Figures 3 and 4. The higher the level of noising, the better a model is able to transfer from one sentiment to another but the less faithful the translation is to the content of the original sentence. The existence of this trade-off is consistent with previous sentiment transfer research [29, 13]. With rationale-based noising BART achieves a better trade-off than the Seq2Seq baseline on the Yelp dataset. The performance of saliency-based noising is similar to that of rationale-based noising with a similar BLEU score on both models, although on Seq2Seq saliency noising is slightly preferable.

On the Amazon dataset the BART model's trade-off across different noising levels is almost identical to that of the Seq2Seq model. BART's high performance with saliency noising is an outlier here compared to the rationale-based models. The combination of BART's ability to generate more fluent sentences and the saliency-based method for masking or removing n-grams of up to four tokens may explain the better performance of that configuration. The results show that all of our models obtain relatively low sentiment transfer accuracies compared

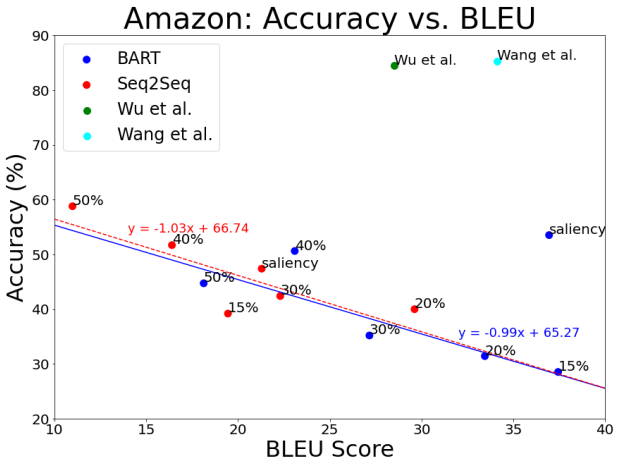

Fig. 4: Test set sentiment transfer vs. BLEU score for BART (solid blue line) and Seq2Seq (striped red line) on the Amazon dataset.

Table 2: Human evaluation results (scores scale 1 to 5, higher is better) on the Yelp and Amazon datasets.

	Yelp			**Amazon**		
Model	Content	Sentiment	Fluency	Content	Sentiment	Fluency
BART 30%	-	-	-	3.4	3.4	3.9
BART 40%	3.9	4.4	4.1	-	-	-
BART 50%	3.2	4.5	4.1	2.8	3.8	3.9
Seq2Seq Saliency	3.4	3.9	3.4	3.3	3.2	3.5
Wu et al. (2019)	4.0	4.4	4.2	4.1	4.0	4.0
Wang et al. (2019)	3.5	3.6	3.8	4.2	4.0	4.1

to the state-of-the-art, suggesting that the Amazon dataset may be less suitable to sentiment transfer based on identifying individual sentiment words.

On Yelp our best model's performance is close to that of state-of-the-art approaches, although both [29] and [27] obtain a slightly better trade-off between accuracy and BLEU. This is shown by these models having accuracy and BLEU scores that lie above the average trade-off, represented by the solid blue line in 3. The simpler structure of examples in the Yelp dataset makes it particularly suitable for the word-based "Mask and Infill" approach [29].

6.2 Human Evaluation

Due to the limitations of automatic evaluation in text generation tasks such as sentiment and style transfer, we performed a human evaluation, broadly following the methodology of [13] and subsequent work on style transfer. Amazon

Mechanical Turk crowd-workers were asked to rate a sample of 100 sentiment transferred generations for each of the models included in the evaluation. The evaluation assesses the fluency, sentiment transfer accuracy and content preservation of the generations, each scored using a Likert scale from 1 to 5. Each output was evaluated by three evaluators.

The results are shown in Table 2. We include the rationale-based BART models with the best trade-off between accuracy and BLEU according to the automatic evaluation (40% noising on Yelp and 30% on Amazon) and the highest overall accuracy (50% noising), as well as the baseline Seq2Seq model with saliency noising. The human evaluation results of Wu et al. [29] and Wang et al. [27], as reported in the original papers, are given as an additional comparison.

The results show that the BART-based models clearly outperform the Seq2Seq baseline on both datasets, with the exception of BART 50%'s content preservation. BART 50% obtains higher sentiment transfer accuracy and similar fluency than the BART models with less noising, but lower content preservation. On Yelp and Amazon the BART-based models obtain higher sentiment transfer accuracy than the Seq2Seq baseline on the human evaluation despite similar or lower automatic classification accuracies. Content preservation is broadly in line with the BLEU scores. The increased fluency of the BART models can be attributed to its pretraining compared to the Seq2Seq baseline which is trained from scratch. This may also explain why BART achieves higher BLEU scores than the Seq2Seq model for the same level of noising.

Compared to state-of-the-art models, the BART 40% model performs better on Yelp than [27] and on par or within 0.1 evaluation points of [29]. Our approach therefore obtains state-of-the-art performance despite performing slightly worse on the automatic evaluation. On Amazon the BART 50% model obtains sentiment and fluency scores within 0.2 points of the state-of-the-art (compared to the automatic classifier which showed a much greater gap in sentiment transfer accuracy), but much worse content preservation. BART 30%'s content preservation is higher but still far below that of the state-of-the-art, and at the expense of lower sentiment transfer accuracy.

In spite of these mixed automatic evaluations, the higher human evaluations still show a clear benefit to using the pretrained model for generation. This suggests that fluency ratings may have an inadvertent impact on other human ratings and that more fluent model outputs are preferred by human evaluators. This is the case even if sentiment translation accuracy, as measured by automatic evaluations, does not match that of a less fluent model.

6.3 Qualitative Analysis

A selection of model outputs on the test data, along with the original sentences and reference human sentiment translations, are shown in Tables 3 and 4. In the first and last Yelp examples the Seq2Seq model manages to translate the sentence, but fluency is worse when compared with the source sentence. While the BART 40% model manages to preserve the content of the sentences, the BART 50% and Seq2Seq models are less successful.

Table 3: Original, reference translation and model sentiment transfer output examples on the Yelp test set.

Model	Output
Original	but was very disappointed with what actually arrived.
Reference	was very happy with what arrived.
BART 40% / 30%	everything was very good with what actually arrived.
BART 50%	but was very happy with what they did.
Seq2Seq Saliency	but it was very nice and easy to say that was dessert.
Original	i do not like the size of the dance floor.
Reference	i love the size of this dance floor!
BART 40% / 30%	i also really like the size of the dance floor.
BART 50%	i also really liked the feel of the dance floor.
Seq2Seq Saliency	i like the size of the dance floor.
Original	its quiet and nice people are here.
Reference	nice people are here, but it is too quiet and boring
BART 40% / 30%	its dirty and rude people are here.
BART 50%	its not like they are very busy.
Seq2Seq Saliency	its quiet people are here.

Although the translation of the first Amazon sentence seems to be relatively obvious, the correct translation of the second and third examples appears less obvious. This is predominantly due to the ambiguity of the sentiment of the original sentences in these examples. These examples also show how the sentence content can become corrupted during sentiment transfer, both with the Seq2Seq examples and some of the BART 50% examples.

7 Conclusion

We proposed a self-supervised training pipeline for text style or sentiment transfer. We diverge from previous work by using an interpretable classifier to identify which sentiment words to mask or remove, and by fine-tuning a pretrained encoder-decoder Transformer, first as a DAE and then with self-supervised learning. The outputs of our models are preferred by human evaluators over a non-pretrained baseline with saliency-based sentiment word identification. The BART model also obtains a better trade-off between sentiment transfer accuracy and content preservation according to automatic evaluation. On Yelp the performance of our approach is comparable to the state-of-the-art, although on the Amazon dataset the approach performs less well. Self-supervised learning improves performance over a model that is already pre-trained and fine-tuned as a denoising autoencoder. As future research this method could be extended to text style transfer for low resource languages, given that parallel sentences are not required.

Table 4: Original, reference translation and model sentiment transfer output examples on the Amazon test set.

Model	Output
Original	the cookbook that comes with it is adequate.
Reference	the cookbook that comes with it is terrible.
BART 40% / 30%	the cookbook that comes with it is terrible.
BART 50%	the cookbook that came with it was terrible.
Seq2Seq Saliency	the only good thing i ve used it is with it is adequate.
Original	it s almost like putting the phone into a high end pair of socks.
Reference	it s almost like putting the phone into a low end pair of socks.
BART 40% / 30%	it s not like putting the phone into a high end pair of shoes.
BART 50%	it s not even worth the effort into a single pair of socks.
Seq2Seq Saliency	it s almost like they are not into a high end pair of socks.
Original	so not that great for leaving on at night.
Reference	perfect for night
BART 40% / 30%	so far that works for leaving on at night. great product.
BART 50%	so far that works for me on at work.
Seq2Seq Saliency	so not that great for leaving a timer on.

References

1. Bastings, J., Aziz, W., Titov, I.: Interpretable neural predictions with differentiable binary variables. In: ACL (1). pp. 2963–2977 (2019)
2. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: NAACL-HLT (1). pp. 4171–4186 (2019)
3. He, J., Wang, X., Neubig, G., Berg-Kirkpatrick, T.: A probabilistic formulation of unsupervised text style transfer. In: ICLR (2020)
4. He, R., McAuley, J.: Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In: Proceedings of the 25th international conference on world wide web. pp. 507–517 (2016)
5. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
6. Keskar, N.S., McCann, B., Varshney, L.R., Xiong, C., Socher, R.: Ctrl: A conditional transformer language model for controllable generation (2019)
7. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: ICLR (2014)
8. Krishna, K., Wieting, J., Iyyer, M.: Reformulating unsupervised style transfer as paraphrase generation. In: EMNLP. pp. 737–762 (2020)
9. Kumaraswamy, P.: A generalized probability density function for double-bounded random processes. Journal of hydrology **46**(1-2), 79–88 (1980)
10. Lample, G., Subramanian, S., Smith, E., Denoyer, L., Ranzato, M., Boureau, Y.L.: Multiple-attribute text rewriting. In: ICML (2019), https://openreview.net/forum?id=H1g2NhC5KQ
11. Lei, T., Barzilay, R., Jaakkola, T.S.: Rationalizing neural predictions. In: EMNLP. pp. 107–117 (2016)

12. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: ACL. pp. 7871–7880 (2020)
13. Li, J., Jia, R., He, H., Liang, P.: Delete, retrieve, generate: a simple approach to sentiment and style transfer. In: NAACL-HLT. pp. 1865–1874 (2018)
14. Logeswaran, L., Lee, H., Bengio, S.: Content preserving text generation with attribute controls. In: Advances in Neural Information Processing Systems. pp. 5108–5118 (2018)
15. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
16. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: EMNLP. pp. 1532–1543 (2014)
17. Post, M.: A call for clarity in reporting BLEU scores. In: WMT. pp. 186–191 (2018)
18. Pryzant, R., Richard, D.M., Dass, N., Kurohashi, S., Jurafsky, D., Yang, D.: Automatically neutralizing subjective bias in text. In: AAAI (2020)
19. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding by generative pre-training (2018)
20. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners (2019)
21. Rao, S., Tetreault, J.R.: Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In: NAACL-HLT. pp. 129–140 (2018)
22. Nogueira dos Santos, C., Melnyk, I., Padhi, I.: Fighting offensive language on social media with unsupervised text style transfer. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 189–194. Melbourne, Australia (Jul 2018). https://doi.org/10.18653/v1/P18-2031, https://aclanthology.org/P18-2031
23. Sudhakar, A., Upadhyay, B., Maheswaran, A.: "transforming" delete, retrieve, generate approach for controlled text style transfer. In: EMNLP/IJCNLP (1). pp. 3267–3277 (2019)
24. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: ICML. pp. 3319–3328. PMLR (2017)
25. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Advances in Neural Information Processing Systems. pp. 3104–3112 (2014)
26. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing. pp. 5998–6008 (2017)
27. Wang, K., Hua, H., Wan, X.: Controllable unsupervised text attribute transfer via editing entangled latent representation. In: Advances in Neural Information Processing Systems. pp. 11034–11044 (2019)
28. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T.L., Gugger, S., Drame, M., Lhoest, Q., Rush, A.M.: Transformers: State-of-the-art natural language processing. In: EMNLP (Demos). pp. 38–45 (2020)
29. Wu, X., Zhang, T., Zang, L., Han, J., Hu, S.: Mask and infill: Applying masked language model for sentiment transfer. In: IJCAI. pp. 5271–5277 (7 2019). https://doi.org/10.24963/ijcai.2019/732, https://doi.org/10.24963/ijcai.2019/732