Balancing Effect of Training Dataset Distribution of Multiple Styles for Multi-Style Text Transfer

Debarati Das **David Ma** **Dongyeop Kang**
Department of Computer Science, University of Minnesota
{das00015, maxxx818, dongyeop}@umn.edu

Abstract

Text style transfer is an exciting task within the field of natural language generation that is often plagued by the need for high-quality paired datasets. Furthermore, training a model for multi-attribute text style transfer requires datasets with sufficient support across all combinations of the considered stylistic attributes, adding to the challenges of training a style transfer model. This paper explores the impact of training data input diversity on the quality of the generated text from the multi-style transfer model. We construct a pseudo-parallel dataset by devising heuristics to adjust the style distribution in the training samples. We balance our training dataset using marginal and joint distributions to train our style transfer models. We observe that a balanced dataset produces more effective control effects over multiple styles than an imbalanced or skewed one. Through quantitative analysis, we explore the impact of multiple style distributions in training data on style-transferred output. These findings will better inform the design of style-transfer datasets.

1 Introduction

Multi-style text transfer is a challenging task today with applications such as automatic domain-appropriate, style-conformant writing (Fu et al., 2018) and AI-assisted stylistic language editing. Text style transfer is an intricate task as all language has a specific context, and those contexts influence the attributes of the language (Hovy and Yang, 2021). Text style transfer is challenging because it involves dealing with the aspects of style coupled with the textual content (Hu et al., 2017; Shen et al., 2017; Lample et al., 2018). This domain's other obstacles include the need for parallel corpus (Jhamtani et al., 2017) and quality training data. As the number of style dimensions increases with multi-style text transfer, not only is the requirement of a jointly annotated corpus across all the stylistic dimensions problematic, but the different styles are not necessarily independent.

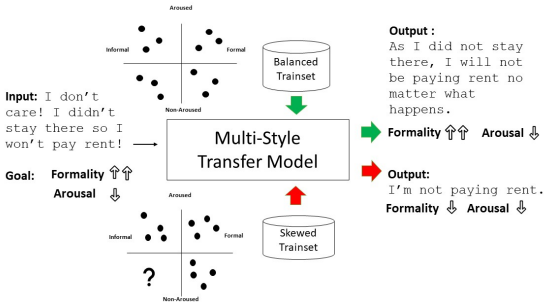

Figure 1: When an input sentence is passed to the multi-style transfer model, to increase formality and decrease arousal, we hypothesize that when the model is trained on a *balanced* joint distribution of formality and arousal (all four style combinations have a 25% representation) - the style transfer is more successful as opposed to when the model is trained on a *skewed* joint distribution (there is no representation of the "informal unaroused" style combination) of styles in the training data.

While "style" can also refer to authorial or domain-specific style, in this paper, we focus on "micro-styles" as defined by (Kang and Hovy, 2021) where they define "micro-style" as a complex combination of different factors such as formality markers, emotions, and metaphors. People intentionally (Troiano et al., 2021) tune these styles in writing differently based on their mood, the person they are addressing, the content of the message, or the platform. Multiple micro-styles can jointly describe a text; for example, a given text could simultaneously be formal and sad. Micro-styles also more easily lend themselves to being represented as spectra with varying degrees of intensity. These points align with our vision of an application where users can edit micro-style aspects of their writing.

Much research exists on models implementing multi-style text transfer and interdependency of micro-styles (Kang and Hovy, 2019; Goyal et al.,

Findings of the Association for Computational Linguistics: ACL 2023, pages 3932–3943

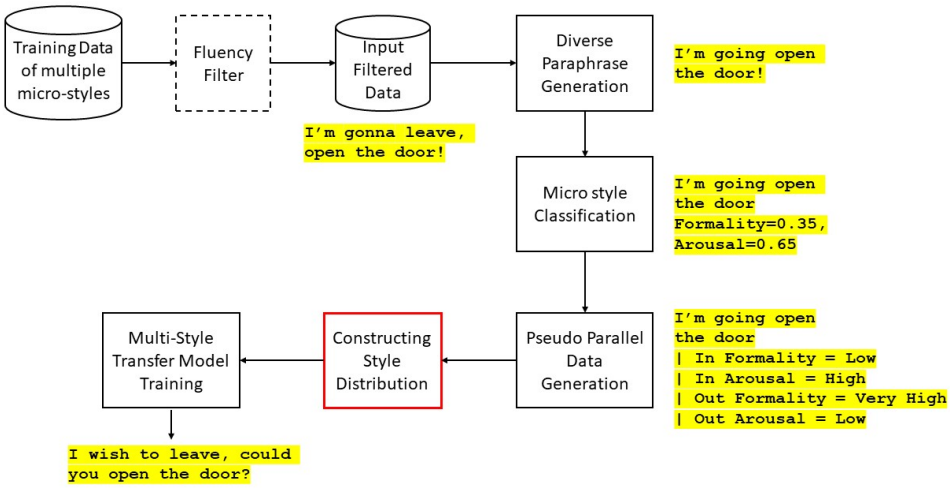

Figure 2: The input sentence transitions through every step in our multi-style text style transfer pipeline. The box in red indicates our main contribution to the pipeline, which helps us explore the effects of joint micro-style combinations on style-transferred output.

2020; Subramanian et al., 2018). However, there needs to be more exploration of the joint distribution of inherent micro-styles in the style transfer training dataset and how these micro-style distributions are related. Therefore, we pose a question - *Can a dataset with minimal variance across multiple micro-style combinations, such that it experiences a "balancing effect", lead to a better style transferred output ?* Figure 1 illustrates our intuition that a dataset that experiences a "balancing effect" will have more control over the multi-style transferred output than a "skewed" dataset. Suppose the style transfer model sees examples of every style combination that can exist - this could aid in the style generation of even unlikely combinations of styles compared to a skewed distribution of these joint micro-styles.

In this research, we consider a multi-style text style transfer pipeline assuming that the user has no access to parallel data or the style of the original text that he wishes to transfer, as would seem natural for a style language editing application. We introduce the changing of the training dataset micro-style joint distributions in such a pipeline and quantitatively explore the impact of this modification on the style transferred output. We perform a set of empirical analyses to demonstrate the influence of joint distributions on style-transferred output and show how this trend varies as the number of micro-styles considered changes. The 'balancing effect' on a training dataset leads to style transferred sentences from even the joint style combinations that are typically rare ("informal unbiased and unaroused"). Our study is the first of its kind on the distribution of micro styles in training datasets for multi-style text style transfer and is likely to have implications for designing datasets for multi-style transfer model training and fall within the context of and align with recent work on characterizing datasets and factors impacting style transfer (Bender and Friedman, 2018; Schoch et al., 2021; Li et al., 2019; Zhang et al., 2020; Gururangan et al., 2018).

2 Multi Style Transfer Pipeline

Datasets: We chose four micro-styles from the style hierarchy defined in Troiano et al.: Formality, Arousal, Sentiment, and Bias, for our study and used publicly available NLP datasets built by other researchers (Rao and Tetreault, 2018; Buechel and Hahn, 2022; Go et al., 2009; Pryzant et al., 2020; Kang and Hovy, 2019) to develop and test our models. Appendix A mentions the details of the datasets and their usage.

Pipeline Overview: Our experimental setup for multi-style transfer is inspired by the work of (Krishna et al., 2020). Like them, we first generate a "diverse" paraphrase of the input sentence, and then the paraphrased sentence is rewritten in the style of choice. Towards this end, we train a paraphrasing model (separately on a parallel paraphrase dataset). Then, the trained paraphrase model is used to create "pseudo-gold" parallel data for training style models.

First, we adopted a pre-trained T5 model (Raffel et al., 2020) to generate paraphrases. This model

was trained for the task of paraphrase generation on the ParaNMT-filtered dataset provided by (Krishna et al., 2020). Once we had this trained paraphrase model, we used diverse beam search (Vijayakumar et al., 2016) to generate diverse fluent paraphrased outputs. An important assumption is that the paraphrase is stripped of its original style and does not leak into the training.

We address this potential issue by training classifiers (Sanh et al., 2019) to predict style on the original and paraphrased datasets and find that all our micro-style classifiers have a classification accuracy of higher than 80% F1, which is acceptable for pseudo-label creation. After we generate diverse paraphrases, we choose the most diverse paraphrase and then derive micro-style classifications for the paraphrased sentence using our trained micro-style classifiers. Therefore each sentence is assigned a classification score for each micro-style label and can form a "pseudo parallel" dataset for training the T5-based joint transfer model. Thus, our approach does not need a parallel dataset.

We then converted the classifier predictions into buckets of style (ranging from "very low" to "very high") based on the chosen style of the original and then paraphrased sentences. The bucketing process is described in Appendix B. After this step, we introduce our contribution of "constructing style distributions" into the pipeline, as illustrated in Figure 2. Following that, we perform multi-style text style transfer. We appended the "bucket" information to the paraphrased sentence to achieve the necessary intensity transfers, as motivated by the original T5 paper (Raffel et al., 2020). We train T5-based style transfer models, where the paraphrased sentence and its style buckets are used as input parameters, while the style buckets assigned to the anchor sentence are used as proxy levels of output style transfer. All model-specific details are provided in Appendix B. For generating sentences from our trained models, we used beam search (Vijayakumar et al., 2016) and nucleus sampling (Holtzman et al., 2019) and chose the top 3 sentences from the generations. The following is an example of the input to the joint style transfer model and the expected output.

`Goal` - Highly increase the formality of the sentence, slightly increase the arousal of the sentence
`Input` - transfer: I'm sad you're going | input formality: low | input arousal: low | output formality: high | output arousal: mid
`Output` - I am sorry you are going to go.

Thus, we implemented a multi-style transfer pipeline to test our hypothesis without any finicky modeling paradigms popular in style transfer research, such as variational inference or autoregressive sampling (He et al., 2020; Subramanian et al., 2018).

Style Combination	Balanced	Skewed
Formal Aroused	3395	8685
Formal Unaroused	3395	2792
Informal Aroused	3395	1275
Informal Unaroused	3395	828

Table 1: Training data statistics (number of samples) for the balanced and skewed settings, when considering the micro-styles of Formality and Arousal.

Constructing Micro-style Distributions We define a "style combination" as a possible combination of the states that the micro-styles can take together - such as 'informal biased negative.' Since there are three micro-styles, each having binary states, the total possible number of style combinations, in this case, is given by $N_c = 2 \times 2 \times 2 = 2^3$. Therefore to generalize, if $|m_i|$ indicates the cardinality of each micro-style and n indicates the number of micro-styles considered, the total possible number of style combinations (N_c) possible is given by :

$$N_c = \prod_{i=1}^{n} |m_i| \quad (1)$$

To create the **balanced** joint distribution of styles, we ensure the standard deviation across the style combinations is close to 0. We do this by down-sampling each style combination, such that the number of samples in each style combination is the same as the least represented style combination. As we increase micro-styles, some micro-style combinations do not occur naturally together, so their representation is close to 0. In such cases, we assume that the least represented style combination is at least 5% of the total dataset. To ensure our comparison across the "balanced" and "skew" settings is fair, we construct a **skewed** dataset with a total sample size that is the same as that of the balanced dataset. Thus, the balanced dataset has a uniform distribution, while the skewed dataset

has a non-uniform distribution. Table 1 shows the number of samples in each style combination of Formality and Arousal, given a "balanced" and "skewed" setting.

3 Experimental Results and Discussion

Evaluation Metrics: Style transfer accuracy metrics quantify how nicely output texts match the desired style. However, more than this metric is required. Motivated by Jin et al., we evaluate style transfer across the three main properties of text style transfer: style transfer accuracy, content preservation, and fluency. We use our custom joint sequence classification models, implemented using HuggingFace libraries (Wolf et al., 2020) to evaluate the style transfer success ratio. Our definition for the Style Transfer Success S_c is the total number of matches between intended and transferred style buckets, divided by the total number of samples. To judge content preserved in style transferred text, we use three metrics: BLEU (Papineni et al., 2002), embedding-based similarity (Wieting et al., 2019) using cosine similarity of two sentence embeddings (Reimers and Gurevych, 2019), and Word Mover's Distance (WMD) (Mir et al., 2019). For fluency, we use measures like perplexity using GPT2 (Radford et al., 2019) and an adversarial classifier using the cross-aligned autoencoder model (Mir et al., 2019).

Experimental Setup: In this paper, we illustrate different micro-style combinations in the training data, for a randomly selected case, with each combination in both the "balanced" and "skewed " settings. Therefore, we consider 6 cases respectively: 1) Formality and Arousal in a balanced setting (`FA balanced`) 2) Formality and Arousal in a skewed setting (`FA skewed`) 3) Formality, Arousal and Bias in a balanced setting (`FAB balanced`) 4) Formality, Arousal and Bias in skewed setting (`FAB skewed`) 5) Formality, Arousal, Bias and Sentiment in the balanced setting (`FABS balanced`) 6) Formality, Arousal, Bias and Sentiment in skewed setting (`FABS skewed`). We construct the training data with the appropriate settings and then pass them through our experimental pipeline (illustrated in Figure 2) and quantitatively evaluate the style transfer results.

Discussion: Table 2 shows examples of style-transferred sentences, given a style-transfer goal from our experimental pipeline for both balanced and skewed settings. E.g., given the objective is to decrease Formality but increase arousal, the sentence " Did you hear about the soldier with 8 limbs? He was army" transforms to "He's an army soldier with 8 legs?". Here, the contraction "He's" indicates a formality decrease, and the replacement of limbs with legs indicates a decrease. The overall arousal of this sentence is higher when it transforms into a question.

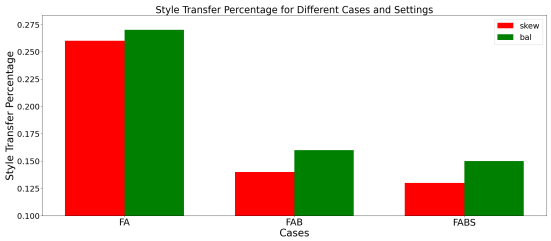

Figure 3: Balancing micro-style distributions leads to a higher multi-style transfer percentage than in the Skewed setting in all the cases.

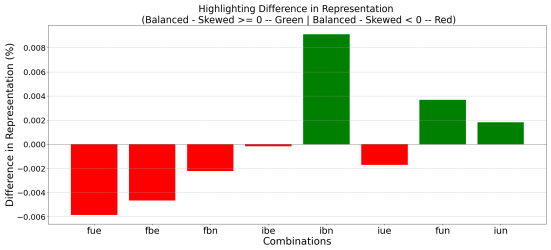

Figure 4: Considering the micro-style combinations such that, Formality [formal = f, informal = i], Bias [biased = b, unbiased = u], and Arousal [aroused = e, un-aroused = n], we observe that the micro-style combinations that are rarer (e.g., informal unbiased neutral (iun)) have more representation in the "balanced" setting than the "skewed" setting.

Figure 3 illustrates that the *balanced setup always has a higher success percentage of style transfer (S_c) than the skewed setup*. We cannot compare the success percentage across cases because matching the exact target and transferred style buckets becomes difficult as the number of micro-styles increases. We can also observe through Table 2 that the *quality of the balanced transferred text aligns better with the style transfer goal than the skewed transferred text.*

In Figure 4, we compare the difference in representation percentage of specific style combinations in the test sample for a specific case where we consider Formality, Arousal, and Bias micro-styles. We observe that a *balanced joint distribution leads to more representation in the style combinations that are less likely to occur.* This is further accentuated as micro-styles increase, as reported in

Style Transfer Goal	Input Text	Balanced Transferred Text	Skewed Transferred Text
↑ Formality ↑ Arousal	Wouldn't it be great if Trump went 3rd party and sucked away millions of Republican votes lol	It wouldn't be nice if Trump went to the third party and swooped millions of Republican votes.	Would it not be great if Trump went 3rd party and sucked away millions of Republican votes?
↑ Formality ↓ Arousal	I didn't know what happiness was until I got married. But by then it was too late.	Until I got married, I didn't even know what happiness was.	I did not know what happiness was till I got married and it was too late.
↓ Formality ↑ Arousal	Did you hear about the soldier with 8 limbs? He was army	He's an army soldier with 8 legs?	Did you hear about the soldier with 8 limbs in the army?
↓ Formality ↓ Arousal	Yeah, I don't understand all the hate.	Yeah I'm not gonna understand the hate.	Yeah I do not understand all the hate.

Table 2: The table shows the style transferred sentences, given an input sentence and the intended style transfer goal, for both the balanced setting as well as the skewed setting.

Appendix C. In Figure 4, we see that rarer style combinations [ibn, fun, iun] show more representation in the balanced case as compared to the skewed case. This supports our intuition that the style transfer model benefits from learning the representation of all possible style combinations that can occur together.

When we consider Formality, Arousal, and Bias micro styles together, the most represented category (30% of samples) is "formal unbiased aroused" (fue). The least represented category (as unlikely to occur together) is "informal unbiased unaroused" (iun) with 1%. We observe that the quantitative evaluation metrics are quite indicative when compared across style combinations. For instance, in Table 3, we observe that *perplexity increases in categories that are unlikely to occur together* (iun). This indicates that the style transfer model is confused by the style distributions present for this style combination.

We do not claim that our method of balancing multiple styles will work even for entangled micro-style combinations, as that is out of the scope of the current paper. However, balancing considerably affects the multi-style transfer output for the range of micro-style combinations we considered, and that has an application in many NLP tasks. This result could hugely influence future studies exploring better ways to balance even the entangled micro-styles.

Setting	Styles	Perp	Adv	BLEU	Cos	WMD
Balanced	fue	**115.16**	0.90	0.77	0.92	0.32
	iun*	**598.58**	0.86	0.78	0.92	0.36
Skewed	fue	116.02	0.90	0.78	0.92	0.32
	iun*	650.47	0.82	0.77	0.93	0.37

Table 3: Comparison of the evaluation metrics for the most represented style combination (fue - formal unbiased aroused) vs the least represented style combination (iun* - informal unbiased unaroused). One key observation is that perplexity increases when the style combinations are unlikely to occur together.

4 Conclusion

Multi-style text style transfer is a challenging problem predominantly plagued by the need for jointly annotated high-quality datasets. There is a clear need for more research about the marginal and joint distribution of inherent micro-styles present in the training dataset used for style transfer. Multi-style text-style transfer typically requires access to large, jointly labeled datasets and many computational resources under typical implementations. More importantly, we would not be able to conveniently tweak the input data distributions in other multi-style text style transfer methods.

In this paper, we implement a multi-style transfer pipeline that subverts the requirement of a jointly annotated dataset of multiple styles by constructing a pseudo-parallel dataset to which we introduce our contribution of constructing style distributions. We then use the modified pseudo-parallel datasets for multi-style transfer. Our modified pipeline effectively allows us to understand the importance of the joint distribution of micro styles in training data and is a substantial contribution.

We quantitatively explore the impact of joint micro-style distributions in the training dataset on the style-transferred output sentences. When the joint micro-style distributions are balanced, there is more control over style-transferred output than with a skewed distribution. These findings will likely inform the design of multi-style transfer datasets and encourage us to explore the micro-style relationships in our datasets.

Limitations

In this research, though we employed automatic evaluation of our multi-style transferred text, we acknowledge that multi-style transfer is challenging to observe with the existing metrics for style transfer evaluation, and human evaluation should

be done as well. As this research paper focuses on exploring the impact of style distributions in the training data on style-transferred output rather than developing a superior multi-style text transfer model, we use quantitative evaluation in this iteration of our paper. We hope that the large sample size and the consistency of applied metrics make our automated approach a reasonable way of evaluating the style transfer output.

This iteration of our paper aims to achieve multi-style transfer across multiple micro styles taken into consideration together as our contribution would aid in constructing a training dataset for multiple micro-style style transfers. We did not explore another exciting question of how balancing multiple micro styles in the training dataset might influence individual style transfer, which could be a promising future direction for our study.

We acknowledge that the classifier's quality sets an upper bound on the best style transfer accuracy that is obtainable. However, the target task is quite complicated without a parallel dataset. Our objective was not to have the most accurate classification of micro styles but to find a means to get acceptable pseudo labels for the micro styles. Individually, all our micro style classifiers had a classification accuracy of 80% F1 and higher, and we deemed this good enough for pseudo-label creation.

We also focused on utilizing the present styles in the training data and classifying them to derive inherent training style distributions instead of dynamically tuning the proportion of styles present in the training dataset. However, tuning these style proportions using techniques such as PPLM (Dathathri et al., 2019) would give us greater control over our experimental pipeline and is an appropriate next step.

Acknowledgement

We thank Vivek Aithal, Priyam Srivastava and Daniel McAndrew for their initial work on the pipeline for multi-style transfer. This was instrumental to our project and helped us get a kickstart on our research.

References

Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Sven Buechel and Udo Hahn. 2022. Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. *arXiv preprint arXiv:2205.01996*.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2019. Plug and play language models: A simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164*.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009.

Navita Goyal, Balaji Vasan Srinivasan, Anandhavelu Natarajan, and Abhilasha Sancheti. 2020. Multi-style transfer with discriminative feedback on disjoint corpus. *arXiv preprint arXiv:2010.11578*.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. *arXiv preprint arXiv:1803.02324*.

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

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

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *The 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. In *International conference on machine learning*, pages 1587–1596. PMLR.

Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Eric Nyberg. 2017. Shakespearizing modern language using copy-enriched sequence-to-sequence models. *arXiv preprint arXiv:1707.01161*.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022. Deep learning for text style transfer: A survey. *Computational Linguistics*, 48(1):155–205.

Dongyeop Kang and Eduard Hovy. 2019. xslue: A benchmark and analysis platform for cross-style language understanding and evaluation. *arXiv preprint arXiv:1911.03663*.

Dongyeop Kang and Eduard Hovy. 2021. Style is not a single variable: Case studies for cross-stylistic language understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2376–2387.

Kalpesh Krishna, John Wieting, and Mohit Iyyer. 2020. Reformulating unsupervised style transfer as paraphrase generation. *arXiv preprint arXiv:2010.05700*.

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

Dianqi Li, Yizhe Zhang, Zhe Gan, Yu Cheng, Chris Brockett, Ming-Ting Sun, and Bill Dolan. 2019. Domain adaptive text style transfer. *arXiv preprint arXiv:1908.09395*.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2020. Automatically neutralizing subjective bias in text. In *Proceedings of the aaai conference on artificial intelligence*, volume 34, pages 480–489.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may i introduce the gyafc dataset: Corpus, benchmarks and metrics for formality style transfer. *arXiv preprint arXiv:1803.06535*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Stephanie Schoch, Wanyu Du, and Yangfeng Ji. 2021. Contextualizing variation in text style transfer datasets. *arXiv preprint arXiv:2108.07871*.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. *Advances in neural information processing systems*, 30.

Sandeep Subramanian, Guillaume Lample, Eric Michael Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2018. Multiple-attribute text style transfer. *arXiv preprint arXiv:1811.00552*.

Enrica Troiano, Aswathy Velutharambath, et al. 2021. From theories on styles to their transfer in text: Bridging the gap with a hierarchical survey. *arXiv preprint arXiv:2110.15871*.

Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45(4):1191–1207.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2019. Neural text generation with unlikelihood training. *arXiv preprint arXiv:1908.04319*.

John Wieting, Taylor Berg-Kirkpatrick, Kevin Gimpel, and Graham Neubig. 2019. Beyond bleu: training neural machine translation with semantic similarity. *arXiv preprint arXiv:1909.06694*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

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

A Dataset Information

We choose four micro-styles from both intended and unintended style categories, based on the style hierarchy as defined in Troiano et al. - Formality, Arousal, Sentiment, and Bias. While formality is considered a "non-targeted intended" micro-style, arousal and sentiment are "targeted intended" micro-styles. We also include subjective bias, an "unintended" micro-style, to ensure we include styles from all hierarchy branches. We built

our micro-style joint classification and style transfer models from multiple publicly available NLP datasets built by other researchers, and we detail these below.

Formality. We use Grammarly's Yahoo Answers Formality Corpus (Rao and Tetreault, 2018), which consists of 105k sentences from two styles: "formal" and "informal" sentences written either in formal or informal modern English. Unlike formal sentences, informal sentences tend to have more misspellings, short forms ("u" instead of "you"), and non-standard usage of punctuation.

Arousal. We use the emotion annotated Emobank dataset (Buechel and Hahn, 2022) based on the three-dimensional VAD model developed by (Warriner et al., 2013). In particular, we transform the Arousal dimension into binary categories such as "arousal" and "non-arousal."

Sentiment. We use the famous Sentiment140 dataset (Go et al., 2009), which consists of automatically annotated tweets, where the tweets containing positive emoticons are assumed as positive. In contrast, those with negative emoticons are assumed to be negative. The training dataset consisted of 1.6M tweets, and the test dataset consisted of 359 tweets. The tweets were preprocessed using NLTK to remove special Twitter-specific symbols like hashtags, usernames, and URLs.

Bias. We use the Wiki Neutrality Corpus by (Pryzant et al., 2020). This is a new parallel corpus of 180,000 biased and neutralized sentence pairs.In order to train our joint classifier models, we used the training dataset from the appropriate micro-style datasets mentioned above. To implement our style distribution hypothesis, we used random samples for training and testing, from the combination of all the dev datasets from the benchmarks corpus by (Kang and Hovy, 2019). This consists of 15 different styles coupled to both content and domain by varying degrees. We wanted to ensure that the dataset used for training our style transfer model and verifying our hypothesis has sufficient indicators of the appropriate micro-styles. This could be done best by using a sample consisting of datasets curated for each individual micro-style (since a jointly annotated dataset with so many styles is not available).

B Multi Style Transfer Pipeline

B.1 Resources used for Training

All models were trained using cloud GPUs on Google Colab Pro and Pro+. We used 1 V100 GPU in its "High-RAM" (52GB) GPU run-time setting to train the paraphrase generation model, while for other models we used 1 P100 GPU at the "standard RAM" setting (32GB).

B.2 Diverse Paraphrase Generation

We adopted a pre-trained T5 model (Raffel et al., 2020), to generate paraphrases. We trained the model on the ParaNMT-filtered dataset provided by (Krishna et al., 2020). This is a subset of the ParaNMT dataset with filters applied to promote lexical diversity, syntactic diversity, and semantic similarity. This model was then used to generate the pseudo-parallel training data for transfer. We selected the t5-small architecture (60 million parameters) as this is approximately 10x smaller than the GPT-2 large model used in (Krishna et al., 2020). We used the hyper-parameters given in Table 4. Based on the recommendation in the appendix of Raffel et al, we used the "paraphrase: " prefix to train the paraphraser model. Once we had this trained paraphrase model, we used diverse beam search (Vijayakumar et al., 2016) to generate diverse paraphrased outputs. The hyper-parameters used for diverse beam search are mentioned in Table 5. We preferred beam search over top-p sampling in order to prioritize fluent paraphrases (Welleck et al., 2019) over unique paraphrases.

`Input` - paraphrase: I love to play my guitar and I do not know why
`Output` - I love playing my guitar and I'm not sure why

Hyperparameters	Value
batch size	8
number of epochs	12
learning rate	1e-4
max sequence length	64

Table 4: Hyper parameters for T5 training for paraphrase generation

Hyperparameters	Value
max length	70
early stopping	True
no repeat ngram size	5
num beams	9
num beam groups	3
diversity penality	0.5

Table 5: Hyperparameters for Beam Search

B.3 Micro-style Classification

We trained a joint sentence classification model to classify the sentence on multiple axes inspired by the approach in Kang and Hovy, which uses an encoder-decoder-based model that learns cross-style patterns with the shared internal representation across styles. Our joint model comprises fully connected layers attached to a DistilBERT model (Sanh et al., 2019), which acts as an encoder. The hyperparameters for this joint model are given in Table 6. This single model effectively replaces the need for a different model for each classification task, significantly reducing the need for computing resources for training and inference. Our joint classifier is essential for downstream tasks like training style transfer models and evaluation. Say we first perform a joint classification of both formality and arousal micro-styles on our datasets, considering we want a multiple-style transfer along the axes of formality and arousal. This results in both formality and arousal pseudo-labels for the sentences. Since these labels are generated algorithmically rather than by hand, we refer to them as *pseudo-labels*. Pseudo-labeled sentences can then be used to generate the pseudo-parallel dataset for training joint style transfer models and directly measure the variation of a style along the axis of interest.

Hyperparameters	Value
train batch size	256
test batch size	512
number of epochs	3
learning rate	1e-4

Table 6: Hyperparameters for Joint Classifier

B.4 Pseudo Parallel Data Generation

We then selected the best paraphrase (most stylistically different from the anchor sentence) based on the cosine distance between the anchor and the paraphrased sentence's style vectors. To enable the transfer model to transfer to specified levels of a particular style, we defined 'very low', 'low', 'mid', 'high', and 'very high' buckets for each micro-style. In the following, we describe the bucket boundaries for our style scores.
`Buckets`: Very Low = [0, 0.2] Low = [0.2, 0.4] Mid = [0.4, 0.6] High = [0.6, 0.95] Very High = [0.95,1]

Using the absolute difference between original text style scores and their best-paraphrased text style scores, we find paraphrasing successfully stripped away both formality and arousal aspects of the text. The same phenomenon has been observed in previous studies, such as (Krishna et al., 2020). To ensure a diverse pseudo-parallel dataset, we retain only anchor-paraphrase pairs that do not match in terms of their style bucket. For example, if an anchor-paraphrase sentence pair is assigned style buckets for formality and arousal, as [very high, low] and [very high, very low], this pair will be retained. However, if both style buckets match, the sentence pair will not be considered diverse enough to remain in the pseudo-parallel dataset. In style transfer models, the paraphrased sentence and its style buckets are used as input parameters, while the style buckets assigned to the anchor sentence are used as proxy levels of output style transfer. The following is an example of the input to the joint style transfer model and the expected output.

`Goal` - Highly increase the formality of the sentence, slightly increase the arousal of the sentence
`Input` - transfer: I'm sad you're going | input formality: low | input arousal: low | output formality: high | output arousal: mid
`Output` - I am sorry you are going to go.

Hyperparameters	Value
train batch size	8
test batch size	8
number of epochs	5
learning rate	1e-4

Table 7: Hyperparameters for T5 for Style Transfer

B.5 Style Transfer Training

Our T5 models were trained on pseudo-parallel datasets created and filtered as described earlier.

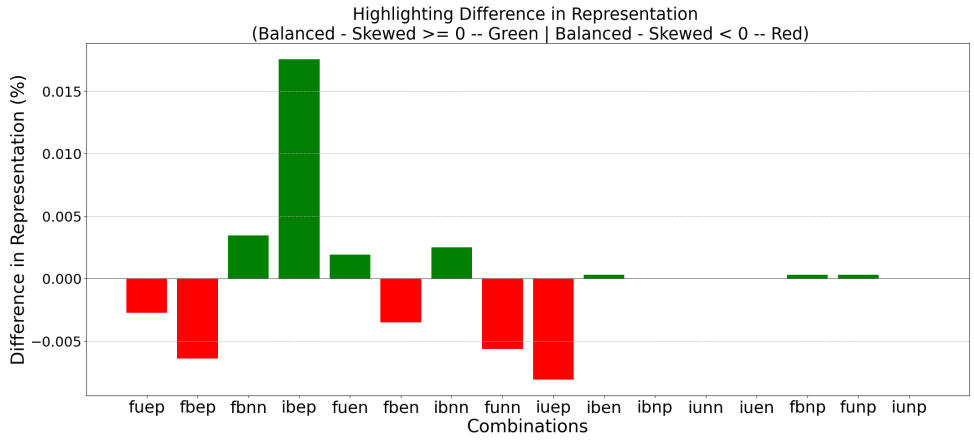

Figure 5: Considering the micro-style combinations such that, Formality [formal = f, informal = i], Bias [biased = b, unbiased = u], Arousal [aroused = e, un-aroused = u], and Sentiment [negative = n, positive = p]; we observe that the micro-style combinations that are rarer have more representation in the "balanced" setting than the "skewed" setting. The categories `fbnp`, `funp` and `iben` have more representation for balanced setting vs skewed setting.

According to the task, we converted the classifier predictions into buckets of style based on the chosen style of the original and then paraphrased sentences. To achieve the necessary intensity transfers, we appended this information to the paraphrased sentence, as motivated by the original T5 paper (Raffel et al., 2020). Hyperparameters are mentioned in Table 7. For generating sentences from our trained models, we used a combination of both beam search (Vijayakumar et al., 2016) and nucleus sampling (Holtzman et al., 2019) and chose the top 3 sentences from the generations.

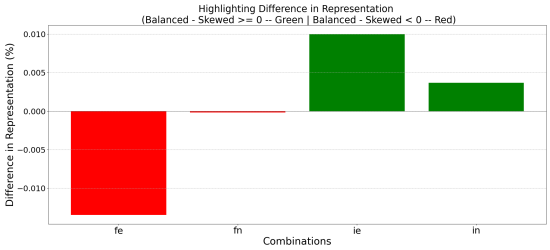

Figure 6: Considering the micro-style combinations such that, Formality [formal = f, informal = i]and Arousal [aroused = e, un-aroused = u]; we observe that the micro-style combinations that are rare (`ie`, `in`) have more representation in the "balanced" setting than the "skewed" setting.

C Some Additional Results

C.1 Impact of Fluency filter on input training data

We find that filtering the original dataset based on fluency metrics always results in better style transferred output as compared to the transferred output when the input dataset is not filtered. This is intuitive, as better quality input prevents confusion in the style transfer model and leads to better quality output. As a result of this finding, we use a fluency filter (adversarial classifier > 0.1 and perplexity < 365), before we conduct any of the rest of our experiments with micro-style distributions.

C.2 Balancing effect on lesser represented style combinations

In Figure 6, we consider the case where we examine Formality [formal = f, informal = i] and Arousal [aroused = e, un-aroused= u] micro-styles and compare the percentage of specific style combinations in the test sample. We observe that as the number of micro styles increases, a balanced joint distribution leads to more representation in combinations that are less likely to occur such as `in` or 'informal and neutral'.

Figure 5 shows a similarly pronounced effect. Here the number of micro styles is increased, and we can observe that the balanced setting shows higher representation than the skewed setting. An example of an unlikely style combination is `fbnp`, or "formal biased neutral and positive". We also observe that as the number of micro-styles increases, there is no representation in some combinations in both settings [`ibnp`, `iunn`, `iuen`, `iunp`]. This is a natural result as some micro-style combinations cannot exist in nature.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
5

☑ A2. Did you discuss any potential risks of your work?
5

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

2,3

☑ B1. Did you cite the creators of artifacts you used?
2,3

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
B

C ☑ Did you run computational experiments?

3,B, C

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
B.1

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
B

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
B

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
A

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.