Preventing author profiling through zero-shot multilingual back-translation

David Ifeoluwa Adelani, Miaoran Zhang, Xiaoyu Shen, Ali Davody, Thomas Kleinbauer, Dietrich Klakow

Preventing Author Profiling through Zero-Shot Multilingual Back-Translation

David Ifeoluwa Adelani, Miaoran Zhang, Xiaoyu Shen, Ali Davody, Thomas Kleinbauer, and Dietrich Klakow
Spoken Language Systems Group, Saarland Informatics Campus, Saarland University, Germany
{didelani,mzhang,xshen,adavody}@lsv.uni-saarland.de
{thomas.kleinbauer,dietrich.klakow}@lsv.uni-saarland.de

Abstract

Documents as short as a single sentence may inadvertently reveal sensitive information about their authors, including e.g. their gender or ethnicity. Style transfer is an effective way of transforming texts in order to remove any information that enables author profiling. However, for a number of current state-of-the-art approaches the improved privacy is accompanied by an undesirable drop in the downstream utility of the transformed data.

In this paper, we propose a simple, zero-shot way to effectively lower the risk of author profiling through multilingual back-translation using off-the-shelf translation models. We compare our models with five representative text style transfer models on three datasets across different domains. Results from both an automatic and a human evaluation show that our approach achieves the best overall performance while requiring no training data. We are able to lower the adversarial prediction of gender and race by up to 22% while retaining 95% of the original utility on downstream tasks.

1 Introduction

Data collections of natural language utterances bear the risk of disclosing sensitive information about the recorded participants, including their gender, race, or political preferences. Unlike explicit mentions of private information, like a user's name or location (Tang et al., 2004; Adelani et al., 2020), such user traits are often encoded rather subtly in a user's speaking or writing style. Nevertheless, they can be predicted with high accuracy by deep learning-based classifiers even when they are not obvious to humans (Elazar and Goldberg, 2018), enabling third-parties with access to the data sets to profile users without their knowledge.

A common method to alleviate this problem is the application of an intermediate transformation step to remove sensitive information via text style transfer. While a number of different style transfer techniques exist (Shen et al., 2017; Fu et al., 2018; Madaan et al., 2020), they require large amounts of text data labeled with user trait information to perform well. Additional annotations need to be provided for every new user trait that the model is expected to handle, multiplying the associated costs and effort. Furthermore, the impact that such transformations can have on the utility of the resulting data is often overlooked. Conversely, we argue that the privacy-utility dichotomy should be at the heart of all research on this topic because it is fairly easy to consider one of the two but difficult to improve both at the same time.

In this paper, we explore a simple yet effective zero-shot text transformation method based on multilingual back-translation. Back-translation (BT) is an alternative approach without the prerequisites of labeled training data. Sensitive user traits can be significantly obfuscated when translated to another language and back (Rabinovich et al., 2017; Prabhumoye et al., 2018) since many concepts cannot easily be mapped across languages. For example, in languages such as Japanese and Korean the speaker's gender can be inferred from the choice of certain pronouns. When back-translating them via an intermediate language that does not make such differences, such as English, these gender indicators will be largely obfuscated.

Results from extensive experiments show that our simple zero-shot text transformer has comparable or even better performance than popular style transfer methods, considering both the privacy and utility of the transformed texts. In summary, we make the following contributions:

1. We propose using multilingual back-translation for hiding users traits. We experiment with using 6 high-resourced languages: German, Spanish, French, Japanese, Russian, and Chinese as the pivot language. This provides more opportunities to pick a language that can hide sensitive information represented in the original language.

Our approach is zero-shot without the need for additional data to train style transfer models.

2. We show that our simple approach is competitive with style transfer models using automatic metrics, and better performance using human evaluation in terms of content preservation and fluency.

3. We perform a comprehensive evaluation on three datasets with popular style transfer methods. These methods have been well studied in the style transfer community, but they have never been evaluated for both privacy and utility preservation in downstream tasks.

2 Related Work

Attribute information such as gender, age, or race are being captured in the deep learning models. Traditional approaches prevented this information leakage via lexical substitution of sensitive words (Reddy and Knight, 2016). In recent years, many text style transfer techniques have been proposed to control certain attributes of generated text (e.g., formality or politeness) while preserving the content. A common paradigm is to disentangle the content and style in the latent space (Shen et al., 2017; John et al., 2019; Cheng et al., 2020). Another stream of work treats text style transfer as an analogy of unsupervised machine translation (Zhang et al., 2018; Lample et al., 2019; Zhao et al., 2019; He et al., 2020) to rephrase a sentence while reducing its stylistic properties (Prabhumoye et al., 2018). Beyond the end-to-end training methods, the prototype-based text editing approach also attracts lot of attention (Li et al., 2018; Sudhakar et al., 2019; Madaan et al., 2020), in which attribute markers of input sentences are deleted and then replaced by target attribute markers. These techniques have been well studied in the text style transfer community, but have never been evaluated for both privacy and utility preservation in downstream tasks. Shetty et al. (2018) and Xu et al. (2019) make use of adversarial training and evaluate on authorship obfuscation. However, they did not include most recent style transfer methods and predictors based on pretrained language models.

3 Multilingual Back-Translation

Problem Scenario In understanding human behaviors and intents, many machine learning applications need to infer important information from users inputs like sentiment, intent, and dialogue act but there is a need to preserve user privacy. We consider a scenario where an adversary attempt to predict demographic attributes of user utterances using a pre-trained attribute classification model. We assume that the adversary already has a pre-trained attribute classification model based on publicly available data. Our goal is to transform the original user input text X to X' such that X' (1) prevents the accurate prediction of user attributes, (2) maintains the utility of downstream NLP tasks, (3) maintains the content of X and (4) is a fluent text itself.

Pivot Language	Translated	Back-translated
DE	Danke Papi	Thank you daddy
FR	merci papi	thank you papi
ZH	谢谢你爸爸	Thank you dad

Table 1: Multilingual back-Translation of "thank u papi" using DE, FR, ZH as pivot languages. User traits can be obfuscated by choosing the proper pivot language.

In this paper, we explore a simple, zero-shot text transformation method through multilingual back-translation. Our assumption is that, as also supported in previous research (Rabinovich et al., 2017; Prabhumoye et al., 2018), text styles can be significantly obfuscated when being translated to another language (pivot language) then translated back. One example is shown in Table 1. The word "papi" is normally used among Latino Americans which exposes their race. When translating them to languages like Chinese then translating back, it becomes the standard form of "dad" and thereby protects the user privacy. Specifically, we define our text transformation function as:

$$X' = T_{L \to en}(T_{en \to L}(X))$$

where L is the pivot language and T is a translation model. We make use of mBART50[1] — an off-the-shelf machine translation model implemented by HuggingFace (Wolf et al., 2020). We consider 6 high-resourced languages as the pivot, so as to ensure a decent quality of machine translation models. The languages chosen are German (DE), Spanish (ES), French (FR), Japanese (JA), Russian (RU), and Chinese (ZH) based on the large amount of resources they have on OPUS (Tiedemann, 2012) and Common Crawl corpora [2].

[1] https://huggingface.co/facebook/mbart-large-50-many-to-many-mmt

[2] https://commoncrawl.org/

Dataset	Attribute Train	Utility Train	Style Train	Dev	Test
DIAL (race)	80K	100K	100K	4K	4K
VerbMobil (gender)	5K	4977	5K	442	1096
Yelp (gender)	2.6M	373K	200K	4K	4K

Table 2: Data splits for DIAL, VerbMobil, and Yelp. The utility task for Yelp and DIAL is sentiment classification while for VerbMobil is dialog act classification.

4 Experiments and Results

4.1 Datasets

In this paper, we conduct experiments on three datasets: DIAL (Blodgett et al., 2016), VerbMobil (Weilhammer et al., 2002) and Yelp (Reddy and Knight, 2016; Shen et al., 2017). These datasets comprise of a variety of domains with either race or gender as the sensitive attribute and they also have annotations for dialog acts and sentiment classification that we use to test the utility of downstream NLP tasks. For **Yelp**, we find two datasets previously used in the style transfer literature, one for gender (YelpGender) (Reddy and Knight, 2016) and the other for sentiment (YelpSentiment) (Shen et al., 2017). The texts are from the same source but each review do not have both gender and sentiment labels. By automatically comparing each review in the test set of YelpGender with the YelpSentiment Dev and Test sets, we created a new **Dev** set and **Test** set with 4K reviews, each with both gender and sentiment information. This can be used for future research to evaluate the utility of Yelp Gender dataset. The dataset is available on Github[3].

Table 2 shows the data splits for three datasets: **Attribute Train**, training set for attribute classification; **Utility Train**, training set for a downstream NLP task; **Style Train**, training set for style transfer, **Dev**, the development set, and the **Test** set. The detailed data description is in Appendix A.

4.2 Experimental Setup

We train five popular style transfer methods: 1) CAE (Shen et al., 2017), (2) BST (Prabhumoye et al., 2018), (3) UNMT (Lample et al., 2019), (4) DLS (He et al., 2020), and (5) Tag&Gen (Madaan et al., 2020). CAE and BST are based on latent representation disentanglement through adversarial training. UNMT and DLS make use of the unsupervised machine translation objective. Tag&Gen is based on prototype-based text editing using frequency ratios method to tag appropriate attribute markers, and generate replacements with a transformer language model. We compare the performance of the style transfer models with multilingual BT models based on mBART50. In addition, we compare with reported results in Xu et al. (2019) on the DIAL dataset. For the attribute, sentiment and dialog act classification, we fine-tune a BERT-base (Devlin et al., 2019) model end-to-end.

Method	Attr. F1↓	Util. F1↑	METEOR↑	GAR↑	P_{Mean} ↑
Original Test set	88.79	75.13	100	48.40	58.69
BT (DE)	81.37	**73.84**	**47.47**	51.83	47.94
BT (ES)	69.44	70.33	32.76	63.50	49.29
BT (FR)	77.72	72.60	41.78	54.88	47.89
BT (JA)	73.77	72.00	34.63	62.22	48.77
BT (RU)	78.81	73.00	42.98	50.38	46.89
BT (ZH)	66.65	71.68	27.61	**80.95**	**53.40**
Adv	65.75	65.70	17.03	_	_
SMDSP	74.85	69.88	28.15	_	_
CAE	35.37	61.63	12.84	22.08	40.30
BST	**13.99**	54.16	5.03	10.60	38.95
UNMT	18.11	64.68	19.95	43.87	52.60
DLS	28.13	66.18	25.04	30.28	48.34
Tag&Gen	44.34	69.74	42.30	23.18	47.72

Table 3: Evaluation on DIAL dataset. Adv and SMDSP result are from (Xu et al., 2019)

4.3 Evaluation tasks and Metrics

Style transfer models are usually evaluated on three tasks: Transfer style (or attribute) strength, content preservation, and fluency (Jin et al., 2021). Although, our desire is for the models to have a very good transfer attribute strength, other evaluation tasks are important since there are several downstream tasks that would benefit immensely from fluency and content preservation. For example, content preservation is critical for question answering systems, and fluency is desirable for dialog generation systems since we may not be able to generate fluent replies with non-fluent inputs.

For **Transfer attribute strength**, we measure the success of the transfer by a drop in attribute F1-score (`Attr`) on the transformed test set. For **Content preservation**, we choose `METEOR` because it takes into account word stems, synonyms and paraphrase leading to better recall. **Fluency** measures grammaticality. Following Krishna et al. (2020), we compute grammaticality acceptance rate (`GAR`) using available fine-tuned models[4] trained

[3] https://github.com/uds-lsv/author-profiling-prevention-BT

[4] https://huggingface.co/textattack/roberta-base-CoLA

Method	VerbMobil Gender					Yelp Gender				
	Attr. F1↓	Utility F1↑	METEOR↑	GAR↑	P_{Mean}↑	Attr. F1↓	Utility F1↑	METEOR↑	GAR↑	P_{Mean}↑
Original Test set	72.24	59.73	100	61.08	49.78	87.92	97.55	100	86.18	73.95
BT (DE)	67.09	**54.19**	**41.21**	68.16	49.12	82.37	**95.45**	**52.42**	88.83	**63.58**
BT (ES)	62.58	50.47	31.38	77.01	49.07	76.51	91.54	38.89	90.37	61.07
BT (FR)	67.98	52.69	36.77	77.28	49.69	76.48	91.80	40.63	91.23	61.80
BT (JA)	65.23	45.52	21.71	87.68	47.42	71.98	92.39	35.47	93.00	62.22
BT (RU)	68.73	52.56	39.27	66.24	47.33	79.11	94.17	45.17	85.88	61.53
BT (ZH)	63.96	51.70	25.80	91.79	**51.33**	72.85	92.22	34.40	**95.57**	62.34
CAE	**49.71**	23.06	6.30	**93.70**	43.32	68.72	88.37	40.18	60.38	55.05
BST	66.51	18.06	1.69	23.18	19.11	51.00	71.00	24.03	58.45	50.62
UNMT	59.97	29.49	13.95	45.99	32.37	68.92	90.71	45.67	77.65	61.28
DLS	61.42	31.66	17.34	47.81	33.85	50.43	82.99	34.55	77.03	61.03
Tag&Gen	69.66	36.40	7.40	67.79	35.48	**33.64**	79.49	36.30	55.32	58.93

Table 4: Evaluation on VerbMobil (low-resource scenario) and Yelp. Comparing style transfer models and BT

Method	Content preservation	Fluency
BT (JA)	4.16	4.76
BT (ZH)	**4.19**	**4.78**
DLS	3.42	4.25
Tag&Gen	2.89	3.42

Table 5: Human evaluation of content preservation and fluency on BT (JA), BT (ZH), DLS, Tag&Gen

Method	sentence
Original	this hotel seems to be very poorly run.
BT (DE)	The hotel seems to be very poorly operated.
BT (JA)	This hotel seems to be very poorly managed.
BT (ZH)	This hotel looks terrible.
CAE	this place is definitely very good.
BST	this hotel seems poorly run.
UNMT	this hotel seems to be very clean .
DLS	i was n't very impressed with this place.
Tag&Gen	this hotel seems to be gorgeous run .

Table 6: Sample sentences for BT (DE), BT (JA), BT (ZH), CAE, BST, DLS, UNMT, Tag&Gen

on CoLA (Warstadt et al., 2019). Lastly, we introduce a new task, **Utility** (`Util`) to measure the performance of the transformed texts on an available downstream NLP task. Further details are in Appendix B. To measure the overall performance across all tasks, we compute an average of all the metrics ($\mathbf{P_{Mean}}$). For transfer attribute strength, we subtract attribute F1 from 100 i.e $(100 - Attr)$ because the value is decreasing while others are increasing. We provide more details in Appendix B.

4.4 Results

Automatic Evaluation We compare the performance of the style transfer models and back-translation models in terms of attribute F1, utility F1, METEOR, and GAR on three datasets (DIAL, VerbMobil and Yelp). Table 3 shows the performance on DIAL dataset. We observe a reduction of $7-22\%$ in attribute F1 by a simple back-translation, with Chinese (ZH) preserving more privacy while maintaining 95% of the original utility and highest score (81%) for fluency. German (DE) has better METEOR score and utility on average but sacrificed a lot of privacy. The BT (ZH) model has similar or better performance as the Adversarial training and SMDSP proposed by (Xu et al., 2019) in privacy preservation, utility and content preservation. However, we find style transfer methods have much better privacy preservation than BT models with $45 - 75\%$ reduction in attribute F1, but they sacrificed a lot in terms of utility, content preservation (< 30 METEOR except Tag&Gen) and fluency ($< 45\%$ GAR), making them not practical for real-life applications.

Table 4 shows the result on VerbMobil dataset. The BT models leads to a reduction of $3.5 - 9.7\%$ in attribute F1 while maintaining over 86% of the original utility F1. We also find them to achieve better performance in METEOR and GAR, although the models are applied in zero-shot settings. The style transfer models performed terribly since they typically require massive amounts of data (Li et al., 2019) and might be skewed in a data-scarcity scenario (5k sentences for VerbMobil). One particular strength of our approach is that it requires no additional data and most suited for zero-shot settings.

We also examined the performance of BT models on Yelp dataset. The style transfer models preserve more gender privacy ($19 - 54\%$) than the BT models ($5-16\%$). However, they have much worse results in terms of utility and fluency. Overall, the $\mathbf{P_{Mean}}$ of BT models is often better than the style transfer models for all datasets.

Human Evaluation We further performed human evaluation for the two best privacy-preserving BT models (ZH and JA) and style transfer models (DLS and Tag&Gen) in terms of content preservation and fluency. We recruited three raters, who are volunteers from our research lab including authors of the paper to evaluate the models. The three volunteers rated 100 sentences per model i.e 400 sentences per rater. The volunteers were not paid for the rating, and were informed that they could in principle, choose to withdraw from the annotation without consequences. We provide the annotation guideline on Github[5].

Table 5 shows the average rating by three professional speakers of English language on 100 sentences in the Yelp dataset, we found out that ZH and JA are rated much higher in content preservation – over 4 (on a $1-5$ Likert scale) while maintaining near perfect fluency (4.7). The inter-agreement Krippendorff α of our human raters is 0.69 for both content preservation and fluency. On the other hand, DLS and Tag&Gen are rated lower on both evaluation tasks. Although, Tag&Gen preserves privacy more on Yelp according to Table 4. Table 6 shows an example sentence confirming the content preservation and fluency of our approach. We provide more examples in Appendix C.

5 Conclusion

In this paper, we propose a zero-shot way to effectively lower the risk of author profiling through multilingual BT using off-the-shelf translation models. We compare our approach with different style transfer models, achieving the best overall performance using an automatic and a human evaluation while requiring no additional training data. In the future, we will (1) analyze how the language choice and translation quality affects the privacy preservation in BT, (2) investigate more on other metrics that can be used to aggregate the the four evaluation metrics corresponding to transfer attribute strength, content preservation, fluency, and utility, and (3) extend the zero-shot BT method with some supervision to improve privacy.

We highlight a few limitations of our work. First, back-translation transformation remove content style but does not necessarily replace attribute markers like style transfer models, for example, given a text "me and my husband ...", style transfer models are more likely to change "husband" to "wife" but back-translation will not. Second, our back-translation technique also inherit some of the problems of machine translation generated texts like hallucination (Raunak et al., 2021). We provide examples highlighting these issues in Appendix C.

[5] https://github.com/uds-lsv/author-profiling-prevention-BT

6 Broader Impact Statement and Ethics

This paper presents an approach to prevent author profiling of sensitive user attributes. We understand there are many ethical concerns around gender and race, however, our definition and evaluation of user traits are constrained by the available datasets we found in the literature. We did not collect any new data to show the strength of our approach. We hope our research helps to protect the profiling of under-represented groups and communities.

Acknowledgements

The presented research has been funded by the European Union's Horizon 2020 research and innovation programme project COMPRISE (http://www.compriseh2020.eu/) under grant agreement No. 3081705. We thank Dana Ruiter for providing the initial draft of the annotation guideline. We also thank members of the Spoken Language Systems Group and anonymous reviewers for their useful feedback on the paper.

References

David Ifeoluwa Adelani, Ali Davody, Thomas Kleinbauer, and Dietrich Klakow. 2020. Privacy guarantees for de-identifying text transformations. In *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pages 4666–4670. ISCA.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas. Association for Computational Linguistics.

Pengyu Cheng, Martin Renqiang Min, Dinghan Shen, Christopher Malon, Yizhe Zhang, Yitong Li, and Lawrence Carin. 2020. Improving disentangled text representation learning with information-theoretic guidance. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7530–7541, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21, Brussels, Belgium. Association for Computational Linguistics.

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

Junxian He, Xinyi Wang, Graham Neubig, and Taylor Berg-Kirkpatrick. 2020. A probabilistic formulation of unsupervised text style transfer. In *International Conference on Learning Representations*.

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

Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. 2019. Disentangled representation learning for non-parallel text style transfer. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 424–434, Florence, Italy. Association for Computational Linguistics.

Kalpesh Krishna, John Wieting, and Mohit Iyyer. 2020. Reformulating unsupervised style transfer as paraphrase generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 737–762, Online. Association for Computational Linguistics.

Guillaume Lample, Sandeep Subramanian, Eric Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2019. Multiple-attribute text rewriting. In *International Conference on Learning Representations*.

Dianqi Li, Yizhe Zhang, Zhe Gan, Yu Cheng, Chris Brockett, Bill Dolan, and Ming-Ting Sun. 2019. Domain adaptive text style transfer. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3304–3313, Hong Kong, China. Association for Computational Linguistics.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1865–1874, New Orleans, Louisiana. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Aman Madaan, Amrith Setlur, Tanmay Parekh, Barnabas Poczos, Graham Neubig, Yiming Yang, Ruslan Salakhutdinov, Alan W Black, and Shrimai Prabhumoye. 2020. Politeness transfer: A tag and generate approach. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1869–1881, Online. Association for Computational Linguistics.

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.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 866–876, Melbourne, Australia. Association for Computational Linguistics.

Ella Rabinovich, Raj Nath Patel, Shachar Mirkin, Lucia Specia, and Shuly Wintner. 2017. Personalized machine translation: Preserving original author traits. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1074–1084, Valencia, Spain. Association for Computational Linguistics.

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. The curious case of hallucinations in neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1172–1183, Online. Association for Computational Linguistics.

Sravana Reddy and Kevin Knight. 2016. Obfuscating gender in social media writing. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 17–26, Austin, Texas. Association for Computational Linguistics.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Rakshith Shetty, Bernt Schiele, and Mario Fritz. 2018. A4nt: Author attribute anonymity by adversarial training of neural machine translation. In *27th USENIX Security Symposium (USENIX Security 18)*, pages 1633–1650, Baltimore, MD. USENIX Association.

Akhilesh Sudhakar, Bhargav Upadhyay, and Arjun Maheswaran. 2019. "transforming" delete, retrieve, generate approach for controlled text style transfer. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3269–3279, Hong Kong, China. Association for Computational Linguistics.

M. J. Tang, Dilek Z. Hakkani-Tür, and AT Gokhan-Tur. 2004. Preserving privacy in spoken language databases. In . *In Proc. of the International Workshop on Privacy and Security Issues in Data Mining, ECML/PKDD*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Karl Weilhammer, Uwe Reichel, and Florian Schiel. 2002. Multi-tier annotations in the verbmobil corpus. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA).

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 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, Online. Association for Computational Linguistics.

Qiongkai Xu, Lizhen Qu, Chenchen Xu, and Ran Cui. 2019. Privacy-aware text rewriting. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 247–257, Tokyo, Japan. Association for Computational Linguistics.

Zhirui Zhang, Shuo Ren, Shujie Liu, Jianyong Wang, Peng Chen, Mu Li, Ming Zhou, and Enhong Chen. 2018. Style transfer as unsupervised machine translation. *arXiv preprint arXiv:1808.07894*.

Yang Zhao, Xiaoyu Shen, Wei Bi, and Akiko Aizawa. 2019. Unsupervised rewriter for multi-sentence compression. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2235–2240.

A Data Description

In this paper, we conduct experiments on three datasets (DIAL, VerbMobil, and Yelp) from Twitter social media, dialog conversations, and business reviews domains. Each of the datasets have either race or gender as the sensitive information, and sentiment classification or dialog act classification as the downstream NLP task to measure utility. Table 2 shows the datasets and their splits: **Attribute Train**, training corpus for the attribute classifier; **Utility Train**, training corpus for an NLP task; **Style Train**, training corpus for style-transfer models, **Dev**, the development set, and the **Test** set.

DIAL created by (Blodgett et al., 2016) for dialectal tweets classification of African American (AAE) and Standard American English (SAE), and each tweet is assigned a predicted race information – AA or White, and sentiment (pos/neg). We make use of the subset of the tweets (Elazar and Goldberg, 2018) with over 80% confidence in race prediction. The final dataset has 180K tweets (90K each for AA and White race), 80K of the tweets are used for training the attribute classifier while the remaining 100K are used for training sentiment classifier and style transfer models.

VerbMobil corpus (Weilhammer et al., 2002) is a dialog corpus of human to human telephone conversation that are scheduling appointments. The English VerbMobil has over 10K utterances, with only 6,538 with gender information and 6,093 with dialog act (DA) information. We make use of 1,096 utterances with both gender information and DA as the test set, and others for training and **Dev**. We used the same training set for attribute classification and style transfer models due to limited data.

Yelp review corpus created by (Reddy and Knight, 2016) has gender annotation (male and female), we combined this dataset with another Yelp review corpus (Shen et al., 2017) with only sentiment annotation. By automatically comparing the reviews in the two datasets, we created a **Dev**

and **Test** set with 4K reviews each with both gender and sentiment information. This can be used for future research to evaluate the utility of Yelp Gender dataset.

B Evaluation tasks and Metrics

Style transfer models are usually evaluated on three tasks: Transfer style (or attribute) strength, content preservation, and fluency (Jin et al., 2021).

1. Transfer attribute strength (`Attr`): For a binary attribute, the goal is to generate a sentence of attribute 1 given an initial sentence with attribute 0. We measure the success of the transfer by a drop in attribute F1-score on the transformed test set.

2. Content preservation(`METEOR`): This is measured using automatic metrics like BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005). We choose METEOR because it has better correlation with human than BLEU that is commonly used. Also, it takes into account word stems, synonyms and paraphrase when computing the score leading to better recall. Recently, it has been popularly adopted by the style-transfer community.

3. Fluency(`GAR`): measures grammaticality. In most cases, this is measured using perplexity on the transformed set. However, Krishna et al. (2020) proposed computing the grammaticality score from a classifier trained on Corpus of Linguistic Acceptability (CoLA) (Warstadt et al., 2019) instead of perplexity because it is unbounded and unnatural sentences with common words may have low perplexity. We compute grammaticality acceptance rate (GAR) using available fine-tuned models[6].

4. Utility(`Util`): we introduce a new task to measure the performance of the transformed texts on an available downstream NLP task. For example, DIAL dataset that is popularly used can also be evaluated for sentiment classification (Xu et al., 2019). Here, we also used the F1-score.

[6] https://huggingface.co/textattack/roberta-base-CoLA

To measure the overall performance across all tasks, we compute an average of all the metrics (P_{Mean}), because all the metrics range from 0 to 100. For the transfer strength, we use (100-F1) since the value is decreasing. Specifically, we compute:

$$P_{Mean} = \frac{100 - Attr + Util + METEOR + GAR}{4}$$

C More Examples:

We provide more examples from the three datasets we considered: Yelp, VerbMobil and DIAL

Method	**sentence**
Original	me and my husband love tokyo lobby !
BT (DE)	me and my husband love Tokyo Lobby!
BT (ES)	I and my husband love Tokyo Lobby!
BT (FR)	I love the Tokyo lobby with my husband!
BT (JA)	my husband and i love the Tokyo lobby!
BT (RU)	I and my husband love the tokyo lobby!
BT (ZH)	My husband and I love Tokyo's amusement park!
CAE	me and my wife loves in san lobby !
BST	my wife and i love the tokyo lobby.
UNMT	me and my wife love the interior !
DLS	it and my wife love this lobby !
Tag&Gen	me and my husband earned tokyo lobby !

Table 7: Yelp: Sample sentences for BT and style transfer models

Method	**sentence**
Original	Lord , i hope this aint nobody i know !
BT (DE)	Sir, I hope this is no one I know!
BT (ES)	Lord, I hope that this is not anyone who knows!
BT (FR)	Lord, I hope this is not someone I know!
BT (JA)	God, I wish there was no one I knew!
BT (RU)	God, I hope it's no one I know!
BT (ZH)	Oh, my God, I wish this wasn't someone I knew!
CAE	<unk> is so good
BST	ENTITY hope not someone I know!
UNMT	Dear God , i hope this is not good ! I miss you
DLS	Lord I hope this would be pretty much ! ENTITY
Tag&Gen	Lord , i hope this is actually know

Table 8: DIAL: Sample sentences for BT and style transfer models

Method	sentence
Original	that gives us plenty of time to chill out before the morning
BT (DE)	this gives us plenty of time to cool off before the morning
BT (ES)	The Committee recommends that the State party ...
BT (FR)	which gives us a lot of time to cool down before the morning
BT (JA)	it gives us enough time to cool down in the morning.
BT (RU)	that gives us plenty of time to rest until the morning
BT (ZH)	So we can have a good rest before the morning
CAE	that is good for me , I am going to be out of town , I am out of town
BST	okay , that is fine , I am going to be out of town , I am out of town
UNMT	okay what do you say we meet on Monday , around two P M
DLS	okay what time what time will you have to go in Trier
Tag&Gen	yeah that is fine for me how about the twenty seventh or twenty seventh or twenty seventh

Table 9: VerbMobil: Sample sentences for BT and style transfer models