Enhancing Text Style Transfer with CycleGAN: Data Augmentation and an Adapted StarGAN for Multi-Style Transfer

Matteo Tomatis
Politecnico di Torino
Turin, Italy
s334271@studenti.polito.it

Giuseppe Priolo
Politecnico di Torino
Turin, Italy
s333027@studenti.polito.it

Federico D'Agostino
Politecnico di Torino
Turin, Italy
s309730@studenti.polito.it

***Index Terms*—GitHub Repository.**

***Abstract*—This paper presents enhancements to text style transfer by building on a CycleGAN-based framework and introducing two key extensions aimed at overcoming common challenges in the field. The first extension focuses on augmenting the training data with synthetic samples generated by large language models. By integrating synthetic reviews with authentic data, our approach seeks to mitigate the effects of data scarcity and reduce overfitting, which is crucial when working with limited resources. The second extension adapts StarGAN—originally developed for image applications—to support multi-style transfer in text. Although our experimental results indicate that the adapted StarGAN model did not perform as well as expected on our current test datasets, it demonstrates promising potential for handling multi-style scenarios. The integration of unified generator and discriminator components with style embeddings lays a solid foundation for future improvements. Both extensions are designed to enhance the model's generalization capabilities and improve its performance in text style transfer tasks, such as altering sentiment or formality without compromising the underlying meaning. Overall, while the synthetic data augmentation has shown to compensate effectively for reduced dataset sizes, the adapted StarGAN approach, despite its initial shortcomings, highlights a promising direction for scalable and flexible multi-style transfer. This work contributes to the field by addressing practical limitations related to data scarcity and computational constraints, and by paving the way for further research to optimize multi-style transfer models.**

I. Introduction

Language is influenced by context, audience, and intent. While humans naturally adapt their language, Artificial Intelligence systems often struggle to manage stylistic nuances. Consequently, developing Natural Language Processing (NLP) tools that control text style is increasingly important.

Text Style Transfer (TST) involves modifying the stylistic attributes of a text (e.g., converting an informal tone to a formal one) while preserving its original meaning. Applications include content moderation and intelligent writing assistance. This paper builds on the CycleGAN architecture proposed in [1] by introducing two key extensions:

- **Data Augmentation:** Expanding the training dataset using synthetic reviews.
- **Adapted StarGAN:** An adapted version of StarGAN (originally used for images) to handle more than two styles simultaneously.

II. Extensions to CycleGAN-Based Text Style Transfer

A. First Extension: Data Augmentation

In [1], the CycleGAN-based text style transfer model was evaluated using the Yelp, GYAFC-music, and GYAFC-family datasets. This extension examines whether augmenting the training dataset improves performance, with a primary focus on the Yelp dataset.

The Yelp dataset comprises restaurant reviews labeled as positive (ratings 4 or 5) or negative (ratings below 3). Due to hardware limitations, training samples were restricted to 40,000 per style. Three dataset configurations were created:

1) **Original Dataset**: Contains only authentic Yelp reviews.
2) **Mixed Dataset**: An equal blend of authentic and synthetically generated reviews.
3) **Synthetic Dataset**: Consists solely of synthetically generated reviews.

B. Second Extension: Adapted StarGAN for Multiple Styles

The original CycleGAN model, as applied in [1], is limited to handling a single pair of styles. The adapted StarGAN model overcomes this by using a single generator and a single discriminator thanks to the incorporation of style embeddings. It is important to note that this StarGAN adaptation is based on the version originally developed for image processing.

B.1 Dataset Extension: For the StarGAN model, three writing styles were considered. Two styles were sourced from the Yelp dataset (with 20,000 samples per style due to hardware limitations), and a third style was derived from a selected portion of the Wikipedia Talk Labels: Aggression dataset [6]. From the Wikipedia dataset, two training corpora were created: one set of approximately 20,000 samples, balanced with respect to the others, and another of 8,000 high-aggression samples.

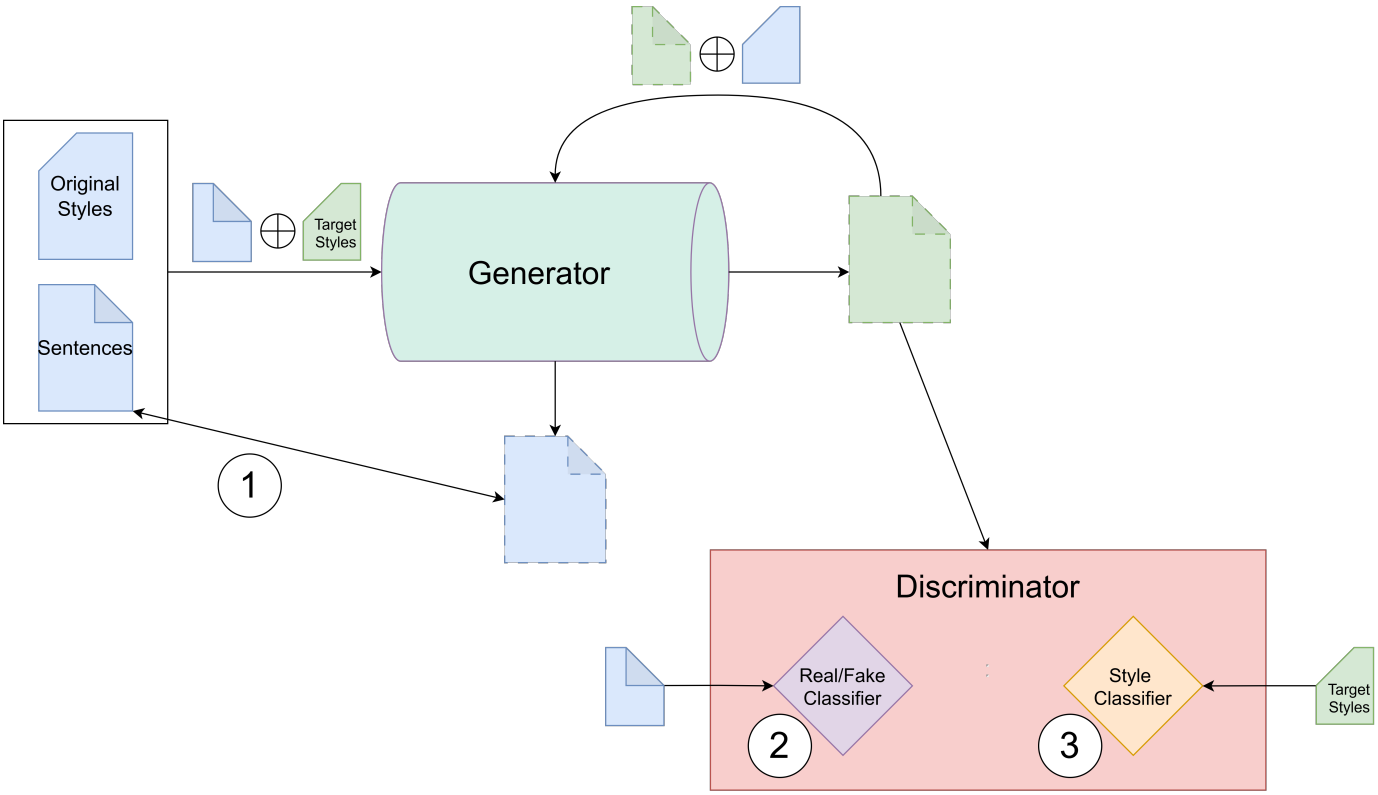

Fig. 1. Illustration of the training process of the StarGAN model. The numbered circles indicate the calculated loss functions: (1) cycle consistency loss, (2) real/fake loss, and (3) style classification loss.

III. Methodology

A. Data Augmentation Strategy

Data augmentation was applied to address data scarcity. The CycleGAN framework used the following components:

- **Generator**: Based on *google-t5/t5-small* for transforming input sentences.
- **Discriminator**: A *distilbert-base-cased* model for distinguishing real from generated text.
- **Style Classifier**: A pretrained model (*ggallipoli/sentiment_classifier_yelp*) to ensure accurate style transfer.

Training was conducted in a self-supervised manner over 10 epochs using a dataset of 80,000 sentences (40,000 per sentiment class), with synthetic reviews generated by ChatGPT 4.0 mini to augment the data.

B. StarGAN Architecture

CycleGAN has proven effective for bidirectional style transfer between two styles. However, extending to multiple styles with CycleGAN requires training separate models for each style pair, which is inefficient. The adapted StarGAN model addresses this by using a single generator and discriminator conditioned on style information via embeddings.

1) Generator: The generator transforms an input sentence x into a target style c while preserving semantic content. The sentence x is embedded and concatenated with a conditional vector representing the target style. To ensure content preservation, the cycle consistency loss is computed as:

$$L_{\text{cyc}} = \mathbb{E}_{x \sim p_{\text{data}}} \left[\| G(G(x, c), c') - x \|_1 \right],$$

enforcing that a round-trip transformation approximates the original sentence.

2) Discriminator and Style Classifier: The discriminator differentiates between real and generated sentences using the adversarial loss:

$$\begin{aligned} L_{\text{adv}} = &-\mathbb{E}_{x \sim p_{\text{data}}} \left[\log D(x) \right] \\ &- \mathbb{E}_{x \sim p_{\text{data}}, c \sim p(c)} \left[\log \left(1 - D(G(x, c)) \right) \right]. \end{aligned} \tag{1}$$

This network is unified with a style classifier, which computes the cross-entropy loss:

$$L_{\text{cls}} = -\mathbb{E}_{x \sim p_{\text{data}}, c \sim p(c)} \left[\log D_c(G(x, c)) \right].$$

This unified approach improves computational efficiency and reduces the overall model size.

3) Objective Function: The total loss for the generator is a weighted sum of the adversarial, cycle consistency, and classification losses:

$$L_{\text{G}} = \lambda_1 L_{\text{adv}} + \lambda_2 L_{\text{cyc}} + \lambda_3 L_{\text{cls}},$$

while the discriminator minimizes:

$$L_{\text{D}} = \lambda_4 \left(L_{\text{real}} + L_{\text{fake}} \right) + \lambda_5 L_{\text{cls}}.$$

A.1 Dataset on Personal Attack: The Wikipedia Talk Labels: Aggression dataset includes comments annotated on a scale from -3 (most aggressive) to 3 (most positive), with several annotations per sentence. This scores were aggregated as their mean, obtaining one label per sentence. Due to differences in sentence length (Yelp reviews average around 9 words versus 69 words for Wikipedia), only comments with 50 words or fewer were retained, reaching a more comparable average lenght of 18 words per sentence. To obtain the balanced set of 20,000 samples, samples labeld -0.5 or less were selected. Alternative experiments with a more strict threshold (only -1.5 or less) yielded about 8,000 samples of a more consistent level of aggression.

IV. Experiments

A. Analysis of Data Augmentation

This section evaluates the impact of data augmentation on CycleGAN-based text style transfer, focusing on two key performance metrics: BERTScore, which assesses semantic fidelity, and F1 score, which measures sentiment classification accuracy.

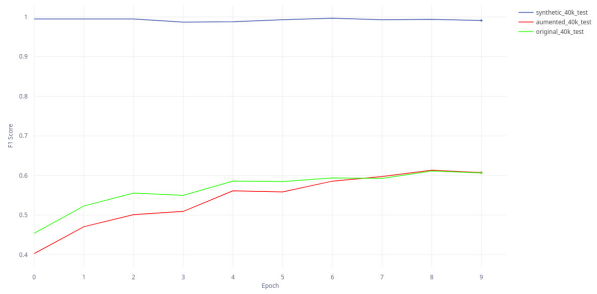

Fig. 2. F1 Score on the test set for Original, Augmented, and Synthetic datasets across 10 epochs.

1) Performance Comparison: Original, Augmented, and Synthetic Datasets: As illustrated in Figure 2, the Synthetic dataset achieves nearly perfect classification accuracy from the very first epoch, significantly outperforming both the Original and Augmented datasets. However, this rapid success comes at the cost of text quality, as seen in later evaluations. Meanwhile, the Augmented and Original datasets start with lower classification performance but steadily improve over the course of training, reaching around 60% accuracy.

Figures 3 and 4 highlight a key limitation of the Synthetic dataset: although it achieves nearly perfect classification accuracy, its BERTScore remains significantly lower than that of the Original and Augmented datasets. This indicates that the synthetic text deviates substantially from natural linguistic patterns, reducing its effectiveness in real-world style transfer tasks. In contrast, the Augmented dataset offers a compromise—while its classification accuracy is lower than that of the Synthetic dataset, it maintains a higher degree of similarity to human-written text, as reflected in its BERTScore.

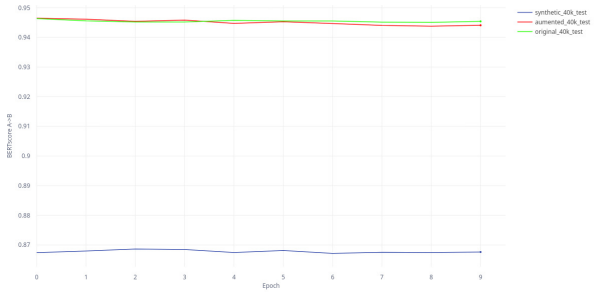

Fig. 3. BERTScore for negative-to-positive style transfer in Original, Augmented, and Synthetic datasets.

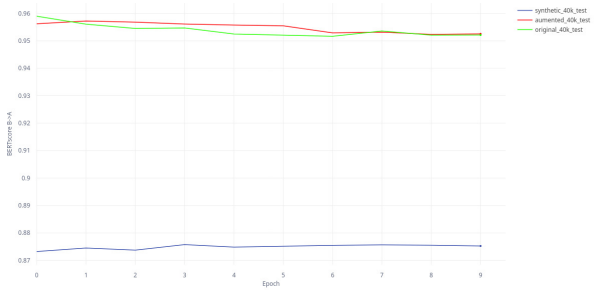

Fig. 4. BERTScore for positive-to-negative style transfer in Original, Augmented, and Synthetic datasets.

Key Observations:

- **Synthetic dataset achieves immediate high classification accuracy**: The model easily distinguishes synthetic samples, but this does not correlate with better semantic quality.
- **Augmented dataset offers a balanced approach**: While classification accuracy is lower than in the Synthetic dataset, it maintains better content fidelity.
- **Trade-off between classification performance and linguistic similarity**: The synthetic text diverges significantly from real Yelp reviews, making it less useful for real-world applications.

2) Impact of Dataset Size and Hardware Constraints: Due to computational constraints, the training dataset was limited to 80,000 sentences—substantially smaller than the 300,000 samples used in [1]—and training was conducted for only 10 epochs, compared to 30 in the original implementation. This reduction in data availability and training duration led to several challenges:

- **Reduced exposure to diverse sentence structures**: The limited dataset size restricts the model's ability to generalize across a broad range of stylistic variations.
- **Increased risk of overfitting**: With fewer training examples, the model may learn spurious correlations rather than robust transformation rules.
- **Compromised fluency in generated text**: The shorter training duration may hinder the model's ability to refine sentence-level coherence and stylistic adaptation.

Data augmentation was employed to mitigate these limitations, yet the overall performance remains constrained relative to prior studies that utilized larger datasets and longer training schedules.

3) Advantages and Limitations of Synthetic Data: While the Synthetic dataset demonstrates near-perfect sentiment classification accuracy, its significantly lower BERTScores highlight a fundamental limitation: the generated text often lacks linguistic richness and diverges from authentic user-generated content.

Key Implications:

- **Simplified sentence structures**: Synthetic data tends to exhibit a more structured and predictable composition, which facilitates classification by the model. However, this characteristic restricts the model's ability to capture and adapt to the nuanced variations present in real-world text.
- **Loss of content fidelity**: The disparity in BERTScore suggests that synthetic reviews do not accurately replicate the linguistic diversity of human-authored content.
- **Potential biases**: Since synthetic data is generated by a pretrained language model, it may introduce biases that distort the model's learning process, leading to undesirable artifacts in style transfer outputs.

B. StarGAN Training Results (Second Extension)

1) Configuration Settings and Metrics: The StarGAN model used a Google T5-small generator and a DistilBERT-based discriminator. Training employed the AdamW optimizer with a learning rate of 5×10^{-5} and a batch size of 64 for 10 epochs. Performance was evaluated using BLEU, ROUGE, and BERTScore across style pairs.

2) Data Configurations: Two data configurations were explored:

1) A balanced dataset of 60k sentences (20k per style).
2) A dataset of 48k sentences with 20k positive, 20k negative, and 8k aggressive samples (introducing class imbalance).

3) Results and Analysis: The StarGAN model yielded low scores across several metrics:

- **BLEU and ROUGE scores** were low, indicating poor lexical overlap and content retention.
- **Moderate BERTScore** suggests some semantic consistency.

Contributing factors include limited dataset size, hardware constraints, potential mode collapse during GAN training, and suboptimal hyperparameter settings.

4) Potential Improvements: Future work could improve performance by:

- Increasing dataset size via further synthetic data generation.
- Utilizing more powerful hardware for larger batch sizes and extended training.
- Pretraining with large language models to enhance initial representations.
- Applying regularization techniques (e.g., gradient penalty, spectral normalization).
- Conducting thorough hyperparameter optimization.

V. Conclusion

This paper presented two complementary extensions to a CycleGAN-based text style transfer framework with the aim of addressing inherent challenges in natural language style transfer. The first extension, which focuses on augmenting the training data with synthetic samples generated by advanced language models, demonstrated its potential in mitigating the limitations imposed by data scarcity and reducing overfitting. Our analysis, shows that while the synthetic dataset quickly achieves high classification accuracy, it often produces text that deviates substantially from the original content, as evidenced by lower BERTScores. Consequently, an augmented approach that mixes synthetic and authentic data strikes a more effective balance between classification performance and content preservation.

The second extension involved adapting the StarGAN architecture, originally developed for image processing, to support multi-style text transfer. By unifying the generator and discriminator through the incorporation of style embeddings, the adapted StarGAN model is theoretically capable of handling more than two styles simultaneously. Although the current implementation did not yield optimal performance on the test datasets, it exhibits promising potential as a foundation for further research into scalable and flexible multi-style transfer solutions.

In summary, our work contributes to the advancement of text style transfer by proposing innovative strategies to overcome issues related to data scarcity and computational constraints.

References

[1] M. La Quatra, G. Gallipoli, L. Cagliero, "Self-supervised Text Style Transfer using Cycle-Consistent Adversarial Networks," Association for Computing Machinery, vol. 15, October 2024.

[2] Jun-Yan Zhu, Taesung Park, Phillip Isola, Alexei A. Efros, "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks," 2017, https://arxiv.org/abs/1703.10593

[3] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, Jaegul Choo, "StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation," 2018, https://arxiv.org/abs/1711.09020

[4] Sudha Rao, Joel Tetreault, "Dear Sir or Madam, May I Introduce the GYAFC Dataset: Corpus, Benchmarks and Metrics for Formality Style Transfer," arXiv preprint arXiv:1803.06535, 2018, https://arxiv.org/abs/1803.06535

[5] Juncen Li, Robin Jia, He He, Percy Liang, "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, June 2018, DOI: 10.18653/v1/N18-1169, https://aclanthology.org/N18-1169/.

[6] Wulczyn, Evan; Thain, Nikhil; Dixon, Lucas, "Wikipedia Talk Labels: Aggression," Figshare, 2017, https://figshare.com/articles/dataset/Wikipedia_Talk_Labels_Aggression/4267550, DOI: 10.6084/m9.figshare.4267550.v1.

VI. Appendix

TABLE I
PERFORMANCE METRICS FOR CYCLEGAN-BASED SENTIMENT TRANSFER (ORIGINAL DATASET)

Metric	Min	Max	Mean	Std Dev	GM	HM
self-BLEU A→B	71.680	78.770	75.225	3.545	75.141	75.058
self-BLEU B→A	61.720	70.950	66.335	4.615	66.174	66.014
ref-BLEU A→B	56.000	57.031	56.515	0.515	56.513	56.510
ref-BLEU B→A	60.910	67.070	63.990	3.080	63.916	63.842
ref-BLEU avg	58.500	62.050	60.275	1.775	60.249	60.223
g-BLEU A→B	63.400	67.020	65.210	1.810	65.185	65.160
g-BLEU B→A	61.320	68.980	65.150	3.830	65.037	64.925
self-ROUGE-1 A→B	0.871	0.918	0.894	0.023	0.894	0.894
self-ROUGE-1 B→A	0.807	0.868	0.838	0.030	0.837	0.836
self-ROUGE-2 A→B	0.790	0.854	0.822	0.032	0.821	0.821
self-ROUGE-2 B→A	0.690	0.771	0.731	0.041	0.730	0.728
self-ROUGE-L A→B	0.871	0.918	0.894	0.024	0.894	0.894
self-ROUGE-L B→A	0.807	0.868	0.838	0.030	0.837	0.836
ref-ROUGE-1 A→B	0.766	0.776	0.771	0.005	0.771	0.771
ref-ROUGE-1 B→A	0.791	0.837	0.814	0.023	0.814	0.813
ref-ROUGE-2 A→B	0.623	0.633	0.628	0.005	0.628	0.628
ref-ROUGE-2 B→A	0.657	0.709	0.683	0.026	0.683	0.682
ref-ROUGE-L A→B	0.762	0.772	0.767	0.005	0.767	0.767
ref-ROUGE-L B→A	0.787	0.833	0.810	0.023	0.809	0.809
BERTScore A→B	0.945	0.946	0.946	0.001	0.946	0.946
BERTScore B→A	0.952	0.959	0.955	0.004	0.955	0.955
Style Accuracy	0.471	0.621	0.546	0.075	0.541	0.536
Style Precision	0.467	0.634	0.551	0.084	0.544	0.538
Style Recall	0.471	0.621	0.546	0.075	0.541	0.536
Style F1 Score	0.454	0.611	0.533	0.079	0.527	0.521

TABLE II
PERFORMANCE METRICS FOR CYCLEGAN-BASED SENTIMENT TRANSFER (AUGMENTED DATASET)

Metric	Min	Max	Mean	Std Dev	GM	HM
self-BLEU A→B	71.696	80.669	76.183	4.487	76.050	75.918
self-BLEU B→A	62.827	73.029	67.928	5.101	67.736	67.545
ref-BLEU A→B	54.860	56.955	55.908	1.048	55.898	55.888
ref-BLEU B→A	61.378	66.006	63.692	2.314	63.650	63.608
ref-BLEU avg	58.119	61.321	59.720	1.601	59.699	59.677
g-BLEU A→B	62.715	67.782	65.249	2.534	65.214	65.180
g-BLEU B→A	62.098	69.112	65.605	3.507	65.518	65.431
self-ROUGE-1 A→B	0.875	0.927	0.901	0.026	0.901	0.900
self-ROUGE-1 B→A	0.812	0.879	0.846	0.034	0.845	0.844
self-ROUGE-2 A→B	0.790	0.870	0.830	0.040	0.829	0.828
self-ROUGE-2 B→A	0.700	0.792	0.746	0.046	0.745	0.744
self-ROUGE-L A→B	0.875	0.926	0.901	0.026	0.900	0.900
self-ROUGE-L B→A	0.812	0.879	0.846	0.034	0.845	0.844
ref-ROUGE-1 A→B	0.763	0.780	0.772	0.009	0.772	0.771
ref-ROUGE-1 B→A	0.792	0.832	0.812	0.020	0.811	0.811
ref-ROUGE-2 A→B	0.613	0.633	0.623	0.010	0.623	0.622
ref-ROUGE-2 B→A	0.656	0.701	0.679	0.023	0.678	0.678
ref-ROUGE-L A→B	0.758	0.775	0.767	0.009	0.767	0.766
ref-ROUGE-L B→A	0.789	0.829	0.809	0.020	0.808	0.808
BERTScore A→B	0.944	0.947	0.946	0.002	0.946	0.946
BERTScore B→A	0.952	0.957	0.955	0.003	0.955	0.955
Style Accuracy	0.416	0.623	0.520	0.104	0.510	0.500
Style Precision	0.408	0.637	0.523	0.115	0.514	0.505
Style Recall	0.416	0.623	0.520	0.104	0.510	0.500
Style F1 Score	0.403	0.613	0.508	0.105	0.498	0.488

TABLE III
PERFORMANCE METRICS FOR CYCLEGAN-BASED SENTIMENT TRANSFER (SYNTHETIC DATASET)

Metric	Min	Max	Mean	Std Dev	GM	HM
self-BLEU A→B	3.264	3.676	3.470	0.138	3.469	3.468
self-BLEU B→A	3.799	4.249	4.043	0.150	4.042	4.041
ref-BLEU A→B	3.971	4.423	4.197	0.157	4.196	4.195
ref-BLEU B→A	4.616	5.104	4.860	0.170	4.859	4.858
ref-BLEU avg	4.293	4.739	4.516	0.158	4.515	4.514
g-BLEU A→B	3.600	4.004	3.802	0.148	3.801	3.800
g-BLEU B→A	4.188	4.657	4.423	0.163	4.422	4.421
self-ROUGE-1 A→B	0.135	0.159	0.147	0.008	0.147	0.147
self-ROUGE-1 B→A	0.152	0.189	0.171	0.012	0.170	0.170
self-ROUGE-2 A→B	0.026	0.046	0.036	0.007	0.035	0.035
self-ROUGE-2 B→A	0.024	0.049	0.036	0.008	0.035	0.035
self-ROUGE-L A→B	0.128	0.152	0.140	0.007	0.140	0.139
self-ROUGE-L B→A	0.141	0.182	0.162	0.014	0.161	0.161
ref-ROUGE-1 A→B	0.182	0.202	0.192	0.007	0.191	0.191
ref-ROUGE-1 B→A	0.201	0.231	0.216	0.010	0.215	0.215
ref-ROUGE-2 A→B	0.035	0.056	0.046	0.007	0.045	0.045
ref-ROUGE-2 B→A	0.037	0.064	0.050	0.008	0.049	0.049
ref-ROUGE-L A→B	0.173	0.194	0.184	0.007	0.183	0.183
ref-ROUGE-L B→A	0.185	0.220	0.203	0.011	0.202	0.202
BERTScore A→B	0.867	0.869	0.868	0.001	0.868	0.868
BERTScore B→A	0.873	0.876	0.874	0.001	0.874	0.874
Style Accuracy	0.987	0.997	0.992	0.003	0.992	0.992
Style Precision	0.987	0.997	0.992	0.003	0.992	0.992
Style Recall	0.987	0.997	0.992	0.003	0.992	0.992
Style F1 Score	0.987	0.997	0.992	0.003	0.992	0.992

TABLE IV
PERFORMANCE METRICS FOR STARGAN-BASED SENTIMENT TRANSFER (YELP + AGGRESSION 60K SAMPLES)

Metric	Avg
self-BLEU	1.686
ref-BLEU	2.349
g-BLEU	1.990
self-ROUGE-1	0.052
self-ROUGE-2	0.003
self-ROUGE-L	0.050
ref-ROUGE-1	0.096
ref-ROUGE-2	0
ref-ROUGE-L	0.096
BERTScore	0.835

TABLE V
PERFORMANCE METRICS FOR STARGAN-BASED SENTIMENT TRANSFER (YELP + SLIM AGGRESSION 48K SAMPLES)

Metric	Last Value
self-BLEU	2.056
ref-BLEU	1.914
g-BLEU	1.983
self-ROUGE-1	0.062
self-ROUGE-2	0.006
self-ROUGE-L	0.059
ref-ROUGE-1	0.109
ref-ROUGE-2	0
ref-ROUGE-L	0.108
BERTScore	0.831