Text Multi Style Transfer using Cycle-Consistent Generative Adversarial Networks

Matteo Bracco
Politecnico di Torino
Turin, Italy
s319845@studenti.polito.it

Francesco La Rota
Politecnico di Torino
Turin, Italy
s323173@studenti.polito.it

Michele Sgrillo
Politecnico di Torino
Turin, Italy
s304046@studenti.polito.it

Abstract—**In this paper, we propose a novel multi-style Generative Adversarial Network (GAN) that supports bidirectional transformations across three distinct styles in an unsupervised setting, eliminating the need for parallel training data. Our method extends a foundational cycle-consistent GAN framework by introducing explicit style tokens and dictionary-enriched tokenizers, which improve style fidelity and reduce token fragmentation. We evaluate our approach on two datasets. The first is a ternary sentiment dataset derived from Amazon Reviews (positive, neutral, negative). The second encompasses Shakespearean, Trump, and poetic lyrics, thus capturing diverse linguistic and stylistic nuances. Experimental results show that our multi-style GAN achieves higher style accuracy and competitive content preservation compared to baseline and plain cycle-consistent approaches, as measured by self-BLEU, self-ROUGE, and comprehensive accuracy-focused metrics.** [1]

I. Introduction

Text Style Transfer [9] has gained significant attention in recent years, with applications spanning sentiment translation, political discourse, and creative writing. While most existing approaches focus on pairwise style transfer (e.g., positive to negative sentiment), real-world scenarios often demand multi-style transformations. Motivated by these needs, we propose a novel multi-style Generative Adversarial Network (GAN) capable of transferring among three distinct textual styles in a bidirectional manner (many to one and one to many), meaning we can specify the input entry and obtain all the outputs or, conversely, give any style in input and obtain the desired output style. We address TST in an unsupervised scenario, i.e., we assume that there is a lack of parallel annotated data to train sequence-to-sequence models.

II. State Of The Art

A Generative Adversarial Network (GAN) [3] consists of two generators and two discriminators. Generators create samples mimicking real data, while discriminators differentiate real from generated samples. This adversarial process improves both components iteratively, improving generation quality.

Text style transfer (TST) has progressed from rule-based methods to deep learning approaches, including variational autoencoders and adversarial networks. Self-supervised techniques, such as the CycleGAN-based method by Gallipoli et al. [6], address the lack of parallel training data by enforcing sequence-level self-supervision.

Traditional CycleGAN-based TST models often struggle with long-range dependencies due to recurrent architectures. To overcome this, Transformer-based [1] generators and discriminators are used, leveraging attention mechanisms [15] for improved encoding and decoding. A pre-trained style classifier further enhances transfer quality by explicitly guiding the transformation, strengthening target style enforcement. On this behalf, DistilBERT [12] is a lightweight Transformer variant that retains 97% of BERT's [2] performance while being 60% faster, making it ideal for real-time inference. DeBERTa [4], on the other hand, improves contextual understanding through disentangled attention and absolute position embeddings, achieving state-of-the-art results on various NLP benchmarks.

III. Methods

Our work builds directly on the foundational code and model presented in Gallipoli et al.'s publication. Motivated by the proposal of expanding their work for more diverse style manipulation, we extend their method to support multi-style transformations by incorporating explicit style tokens that enable flexible, bidirectional conversions across three distinct styles. This research thus serves as both an extension and a deeper exploration of the original framework, aiming to refine and generalize the self-supervised cycle-consistent approach for a broader range of stylistic transformations.

A. Datasets

We prepared two distinct datasets upon which we evaluated the performance of the proposed architecture.

a) Amazon: We started by focusing on sentiment analysis, expanding on Gallipoli's work with the Yelp non-parallel dataset [14]. While we considered the DynaSent dataset [11], its lengthy, context-heavy sentences made preserving semantic integrity during style transfer challenging. Additionally, its skew toward neutral reviews hindered balanced multi-style transformations, making it less suitable for our experiments.

We selected Amazon Reviews [5], constructing a three-style version based on *raw reviews on Food*. This dataset,

[1] **The GitHub repository of the project can be found here:** https://github.com/s304046/nlpproject/

originally over 13 million reviews (1–5 stars), was filtered to exclude texts shorter than 15 or longer than 100 characters. We applied a Star-Based Categorization, labeling 1, 3, and 5-star reviews as Negative, Neutral, and Positive, respectively, discarding others. Further preprocessing included expanding contractions, decoding HTML entities, removing URLs, and filtering non-English content.

We found a strong bias in the neutral class, as 3-star reviews contained both mildly positive and negative sentiments. To address this, we ”neutralized” reviews with a double approach: we selected only sentences with a reported neutrality score (computed by Python library TextBlob) greater than a given threshold, and we also handled extreme adjectives and punctuation: ”love” became ”like”, exclamation marks were removed, etc. This improved both classifier performance and GAN generation across styles. This pipeline left us with 21k balanced examples per class, though truly neutral samples remained rare.

b) Authorship-Stylized Datasets: To introduce greater stylistic diversity, we utilized a dataset of approximately 21k sentences drawn from Lin Daiyu's poetic lyrics (translated from Chinese to English), Trump speeches, and Shakespearean plays [13]. The dataset was imbalanced, with 7k Trump samples and 6.5k for both Lyrics and Shakespeare.

Unlike sentiment transfer, style transfer across these distinct literary and rhetorical forms poses a greater challenge, requiring the model to capture unique syntactic and lexical features. Preprocessing included removing non-English content, managing punctuation, and correcting misspellings across styles. While this dataset was significantly smaller than the Amazon corpus, it still yielded meaningful and insightful results.

B. Metrics

Evaluating style transfer in a non-parallel setting is challenging due to the absence of direct reference sentences. We employ a combination of self-BLEU [10], self-ROUGE [7], and style accuracy.

Self-BLEU measures content preservation by computing BLEU-4 between the generated sentence $\hat{x}$ and the original input x (BLEU-4 evaluates how many 4-grams in $\hat{x}$ appear in x). Higher scores indicate greater lexical similarity.

Self-ROUGE follows a similar principle, using ROUGE-1, ROUGE-2, and ROUGE-L to capture content overlap, emphasizing recall to detect retained information.

Style Accuracy quantifies the effectiveness of transformation using a pretrained classifier, comparing the expected output style of the generated text with the output style of the classifier for that sentence.

To balance content preservation and style accuracy, we introduce composite metrics: **acc-BLEU** (arithmetic mean of self-BLEU and style accuracy), **g-acc-BLEU** (geometric mean), and **h-acc-BLEU** (harmonic mean).

C. Architecture

a) Pretrained Classifiers: In the Amazon experiment, we extended the original binary style classifier to a ternary classification setting to accommodate the three distinct style categories present in our dataset. Specifically, we fine-tuned a DeBERTa-base model on an Amazon-based dataset labeled with positive, negative, and neutral sentiment. Additionally, for the Shakespeare-Trump binary classification task, we trained a DistilBERT model. For the ternary Shakespeare-Trump-Lyrics dataset, we trained both a DistilBERT and a DeBERTa model to evaluate their effectiveness in distinguishing between the three styles.

b) Cycle-GAN: Our GAN framework retains the original cycle-consistent adversarial setup for both extensions, consisting of two generators and two discriminators. However, we adapted its functionality to accommodate three distinct styles simultaneously. Specifically, we modified the generator mappings as follows:

- **Generator** G_{AB}: Converts style A to both styles B and C.
- **Generator** G_{BA}: Converts from C to A and from B to A.

A key modification was the introduction of a *style token*, which explicitly instructs the generator on the desired transformation direction. For example, to convert a positive review into a neutral one, the input is prepended with a token such as `[pos->neu]`. All valid style tokens were incorporated into the tokenizer's vocabulary before training to prevent unintended token fragmentation. Without this explicit token control, we observed that the model frequently collapsed into repetitive outputs.

We decided to aim for a one-to-many/many-to-one model, rather than a many-to-many approach as we wanted to reduce scalability issues. Because training time increases along with the number of involved styles, we propose an approach which could easily be scaled to 4, 5, n styles. Moreover, modules are trained end to end, meaning that provided the input style A, A-BC and BC-A generators are trained simultaneously.

During training, each epoch systematically covers all transformation directions: A to B and B to A, A to C, and C to A. Instead of using a conventional linear learning rate scheduler, we employed *Cosine Annealing with Warm Restarts* [8], allowing the learning rate to oscillate periodically between its initial value and near-zero across epochs. This approach helps stabilize training and improves convergence in multi-style transformations.

IV. Experiments

A. Amazon

To evaluate the performance of the proposed ternary GAN architecture, we conducted two experiments on the Amazon sentiment dataset, leveraging a pre-trained DeBERTa-base ternary classifier to assist in style evaluation. These experiments aimed to assess how effectively the model transfers between different sentiment styles:

- Ternary GAN with DeBERTa-base classifier, Positive to Neutral/Negative
- Ternary GAN with DeBERTa-base classifier, Neutral to Positive/Negative

This setup allowed us to systematically analyze the adaptability of the model across different sentiment shifts, particularly assessing whether the transformation quality varied when the source text contained strong stylistic markers (e.g., Positive vs. Negative) versus when it started from a more neutral baseline.

B. Shakespeare

For the Shakespeare Extension we trained both the standard cycle-consistent adversarial model and the newly proposed ternary architecture to evaluate its scalability across different style transfer tasks. This allowed us to assess how well the multi-style approach extends beyond the sentiment and formality transfer tasks explored in previous studies by Gallipoli et al.

We leveraged the stylistic peculiarities and distinctive lexicon of Shakespearean text by introducing a targeted vocabulary expansion. Rather than adding all new tokens to all tokenizers indiscriminately, we refined our tokenization strategy by augmenting each generator and discriminator with only the relevant lexicon specific to its respective transformation direction. Specifically, we extracted domain-specific tokens from the Shakespeare and Trump corpora that were not originally present in the respective tokenizer vocabularies. These new tokens were identified using a rule-based filtering mechanism that excluded numerical values, punctuation, placeholders, and overly short or long words.

The extracted Shakespearean words were incorporated into the tokenizers for the Trump-to-Shakespeare generator (G_{AB}) and its corresponding discriminator (D_{AB}), ensuring that the model could generate and classify Shakespearean-style text more effectively. Conversely, Trump-specific tokens were added to the Shakespeare-to-Trump generator (G_{BA}) and its corresponding discriminator (D_{BA}), allowing the model to better handle the linguistic features of Trump's speech style. This targeted vocabulary expansion helped mitigate token fragmentation issues and ensured more stylistically coherent text generation.

When moving to the ternary classifier for Shakespeare, we kept the same approach, passing the relevant lexicon specific to its respective transformation direction(s), and we focused on a comparison in the results obtained by the GAN when using different models are classifiers.

Thus, the experiments run over the Shakesperean dataset were:

- Plain and Dictionary-Enriched Binary GAN, Trump to Shakespeare
- Dictionary-Enriched Ternary GAN with DistilBERT and DeBERTa classifiers, Lyrics to Trump, Shakespeare

V. Results

A. Amazon Experiments

1) Classifier: Table I shows that the DeBERTa ternary classifier, validated over a dataset of more than 3k samples (uniformly divided among all three classes), achieved a strong overall accuracy of 86.12%. The model performed best on the positive class (F1-score: 0.94), suggesting that positive sentiment is the easiest to detect due to its distinct linguistic patterns. The neutral class had a balanced performance (F1-score: 0.82), with higher recall (0.86) than precision (0.79), indicating some misclassification into other categories. The negative class, while maintaining high precision (0.86), had the lowest recall (0.78). This asymmetry implies that neutral and negative sentiments may share overlapping linguistic patterns, making separation more challenging.

2) GAN: We trained the proposed GAN ternary architecture with two input directions to compare results involving neutral class, which was the one we had the most issues with. Table II shows the hyper-parameter configuration we kept for both the experiment run over the Amazon Dataset, with the DeBERTa-base Classifier we just trained to help with the style feedback. As a note, we set the number of epochs to 20 as we had limited training resources, but it should be much higher to obtain optimum results, as it scales with the number of styles the model has to learn.

a) Positive to Neutral/Negative: The qualitative examples in Table III show that the model effectively modifies sentiment while maintaining sentence structure. Transformations from positive to negative are strong, with clear shifts in sentiment polarity, while positive-to-neutral transformations appear more conservative, often softening expressions rather than fully neutralizing sentiment.

Table IV highlights the model's performance across different transformations. The self-BLEU and self-ROUGE scores indicate good content preservation, with higher values in B→A directions, suggesting that converting back to positive tends to retain more original phrasing. Style accuracy is significantly higher in the positive-to-negative setting (75.92%) compared to positive-to-neutral (43.02%), indicating that the model finds it easier to generate strongly negative sentences than neutral ones. Achieving neutrality remains challenging, often retaining either positive or negative connotations rather than fully neutralizing the text. This is probably due to the fact that, as stated previously, the neutral samples are more a mix of neg and pos styles, rather than true neutral sentences, due to the nature of 3-star reviews.

b) Neutral to Positive/Negative: Table V shows that transformations from neutral to positive are generally subtle, often enhancing descriptions rather than fully changing sentiment, such as "An average coffee" becoming "An excellent coffee." In contrast, neutral-to-negative transformations tend to introduce stronger polarity shifts, sometimes exaggerating negativity.

The quantitative results in Table VI confirm that neutral-to-negative shifts are more distinct, with a higher style accuracy (50.07%) compared to neutral-to-positive (37.94%). This aligns with the positive-to-negative results in Table IV, where negative generations were more accurate. The self-BLEU and self-ROUGE scores indicate strong content preservation across both directions.

Generally, the model struggles more to amplify positivity than to intensify negativity, exhibiting a tendency to strengthen sentiment polarity more effectively than to dampen it. Again, this is because the neutral training samples, which tend to be more negative than neutral/positive as for the nature of the dataset, skew the model towards negativity.

B. Shakespeare Experiments

1) Classifier: We trained a DistilBERT classifier for both binary and ternary experiments. For the ternary experiment we also trained a DeBERTa classifier to compare their performances over the most challenging style, which proved to be the lyrics class. We validated them both over 1000 samples, divided equally among the three classes (trump class has around 50 more samples than Shakespeare and lyrics).

a) Binary: Table VII shows that the binary DistilBERT classifier achieved an overall accuracy of 91%, performing equally well on both classes: precision and recall are well-balanced, suggesting that the model is not biased toward either class.

b) Ternary: Table VIII compares the performance of DeBERTa and DistilBERT on the ternary classification task. DistilBERT achieved a higher overall accuracy (91%) compared to DeBERTa (86%), which however showed better balance across classes, with more consistent F1-scores across Trump (88%), Lyrics (86%), and Shakespeare (84%), whereas DistilBERT had more variation, performing strongest on Trump (90%) and Shakespeare (87%) but slightly weaker on Lyrics (84%).

The lyrics style posed the greatest challenge, with DeBERTa achieving lower recall (85%) and DistilBERT achieving lower precision (77%).

It is worth noting that the non-enhanced model systematically "Trumpified" every output, causing Shakespearean-specific words to be lost. This severely impacted the distinctiveness of generated Shakespearean text, making results unsuitable for evaluation. For this reason, we only present the enriched results.

2) GAN: Table IX shows Hyper-parameter settings for all Shakespeare GAN experiments, both binary ones (enriched and not) and all the ternary ones. As before, with extended resources the number of epochs could be doubled to obtain even more meaningful results and solid generated phrases.

a) Binary Trump to Shakespeare: We trained gallipoli's original GAN over a different dataset (ternary Shakespeare dataset from which we only took Shakespeare and Trump styles).

Table X presents examples of style transfer using both the baseline (plain) and dictionary-enriched GAN models. The enriched model generates sentences that better align with the target style by incorporating more expressive and stylistically appropriate vocabulary.

Table XI provides a numerical comparison of both models. The plain model outperforms the enriched model in self-BLEU and self-ROUGE scores, suggesting that the plain model retains more content similarity to the original sentence. This is expected, as the enriched model introduces a broader vocabulary, leading to more diverse paraphrasing rather than strict content preservation. However, the style accuracy significantly improves in the enriched model, increasing from 49.03% to 72.14%. This suggests that, despite a reduction in lexical overlap with the original text, the enriched model achieves a more accurate stylistic transformation. The acc-BLEU metric is also higher in the enriched model, reflecting a better balance between content preservation and style adherence.

b) Dictionary-Enriched Lyrics to Trump, Shakespeare: Table XII presents qualitative examples of the style transfer using the DistilBERT-based GAN. The generated sentences demonstrate that the model effectively captures stylistic elements while maintaining coherence. Lyrics-to-Trump transformations produce more structured, assertive language, whereas Lyrics-to-Shakespeare transformations introduce more archaic phrasing. Due to the limited number of training epochs, even though the model improves stylistic alignment, some outputs still lack fluency. Certain generated sentences remain overly literal, indicating that while enriched tokenization aids in stylistic adaptation, fluency remains a challenge.

Table XIII provides a direct comparison between DistilBERT and DeBERTa contribution in the ternary GAN setting. Both models achieve similar self-BLEU and self-ROUGE scores, indicating comparable content retention. However, DistilBERT exhibits better style accuracy, particularly in the Lyrics-to-Shakespeare setting (73.54% vs. 68.00%), suggesting stronger adaptation to Shakespearean language.

DeBERTa achieves slightly better BLEU scores, indicating stronger content consistency (but it could also mean that the model doesn't really change the input). The acc-BLEU and geometric-acc-BLEU scores are consistently higher for DistilBERT in Lyrics-to-Shakespeare but favor DeBERTa in Lyrics-to-Trump, suggesting a trade-off between content preservation and style adherence.

VI. Discussion and Future Work

The results indicate a strong correlation between GAN performance and the accuracy of the style classifier. Specifically, when the classifier struggles to differentiate styles, the GAN also fails to enforce proper style transfer.

One limitation of our approach was dataset size, particularly in the authorship experiment. Future work should explore scaling up the dataset and integrating more complex generator architectures to assess their impact. Additionally, hyper parameter tuning and alternative learning rate schedulers should be investigated to optimize training stability and convergence speed.

Another promising modification could be adding style tokens to the discriminators, allowing them to leverage explicit style cues during training, potentially improving classification and transfer accuracy.

References

[1] Dai, N., Liang, J., Qiu, X., and Huang, X. Style transformer: Unpaired text style transfer without disentangled latent representation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Florence, Italy, July 2019), A. Korhonen, D. Traum, and L. Màrquez, Eds., Association for Computational Linguistics, pp. 5997–6007.

[2] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[3] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks, 2014.

[4] He, P., Liu, X., Gao, J., and Chen, W. Deberta: Decoding-enhanced bert with disentangled attention, 2021.

[5] Hou, Y., Li, J., He, Z., Yan, A., Chen, X., and McAuley, J. Bridging language and items for retrieval and recommendation, 2024.

[6] La Quatra, M., Gallipoli, G., and Cagliero, L. Self-supervised text style transfer using cycle-consistent adversarial networks. *ACM Trans. Intell. Syst. Technol. 15*, 5 (Nov. 2024).

[7] Lin, C.-Y. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out* (Barcelona, Spain, July 2004), Association for Computational Linguistics, pp. 74–81.

[8] Loshchilov, I., and Hutter, F. Sgdr: Stochastic gradient descent with warm restarts, 2017.

[9] Mukherjee, S., and Dušek, O. Text style transfer: An introductory overview, 2024.

[10] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (Philadelphia, Pennsylvania, USA, July 2002), P. Isabelle, E. Charniak, and D. Lin, Eds., Association for Computational Linguistics, pp. 311–318.

[11] Potts, C., Wu, Z., Geiger, A., and Kiela, D. Dynasent: A dynamic benchmark for sentiment analysis, 2020.

[12] Sanh, V., Debut, L., Chaumond, J., and Wolf, T. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020.

[13] Shao, Z., Zhang, J., Li, H., Huang, X., Zhou, C., Wang, Y., Gong, J., Li, C., and Chen, H. Authorship style transfer with inverse transfer data augmentation. *AI Open 5* (2024), 94–103.

[14] Shen, T., Lei, T., Barzilay, R., and Jaakkola, T. Style transfer from non-parallel text by cross-alignment, 2017.

[15] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need, 2023.

Tables

TABLE I
Performance metrics for DeBERTa Ternary Classifier over Amazon sentiment dataset.

Label	**Precision**	**Recall**	**F1-Score**	**Support**
neutral	0.79	0.86	0.82	1102
positive	0.94	0.95	0.94	1102
negative	0.86	0.78	0.82	1101
Overall Accuracy	-	-	**0.8612**	**Tot**: 3305

TABLE II
Hyper-parameter settings for both Amazon GAN experiments

Parameter	**Value**
Generator Model	BART-base
Discriminator Model	DistilBERT-base-cased
Learning Rate	1e-4
Batch Size	8
Number of Epochs	20
Max Seq Length	64

TABLE III
Comparison of Base Sentences and Generated Style Transfers for the Positive to Neutral/Negative experiment.

<table>
<tr><th colspan="3">One to Many (Positive → Negative, Neutral)</th></tr>
<tr><th>Original (Positive)</th><th>Generated (Negative)</th><th>Generated (Neutral)</th></tr>
<tr><td>Amazing product with great nutritional properties! Price is the only questionable factor.</td><td>Awful product with no nutritional information! Price is the only negative factor.</td><td>Arrived with no nutritional label! Price is the only questionable factor.</td></tr>
<tr><td>It was a gift and she loved it.</td><td>It was a recall and she threw it away.</td><td>It was a gift and she did not like it.</td></tr>
<tr><td>Super tasty! Awesome customer service!</td><td>Super salty! Unhappy customer!</td><td>Rip Off! Buyer Beware!</td></tr>
<tr><th colspan="3">Many to One (Neutral, Negative → Positive)</th></tr>
<tr><th colspan="3">Neutral → Positive</th></tr>
<tr><td>It is not what I expected, it is average, it is ok.</td><td colspan="2">It is what I expected, it is fine, it is perfect.</td></tr>
<tr><td>I was looking forward to trying these out, but they did not match my expectations.</td><td colspan="2">I was looking forward to trying these out, but they exceeded my expectations.</td></tr>
<tr><td>Probably will not buy this coffee again. Not quite flavorful.</td><td colspan="2">Definitely will buy this coffee again. Very flavorful.</td></tr>
<tr><th colspan="3">Negative → Positive</th></tr>
<tr><td>Do not buy these. I was excited to find them at such a price, and I am severely disappointed.</td><td colspan="2">Love these. I was excited to find them at such a good price, and I am well satisfied.</td></tr>
<tr><td>The chocolate was so stale. I had to throw both bags out.</td><td colspan="2">The chocolate was so fresh. I had to order both bags again.</td></tr>
<tr><td>Stale and a waste of money.</td><td colspan="2">Delicious and a great price.</td></tr>
</table>

TABLE IV

Validation metrics obtained over both configurations: A Pos|B Neu and A Pos|B Neg.

Metric	A Pos|B Neu	A Pos|B Neg
self-BLEU A→B	43.95%	43.93%
self-BLEU B→A	53.05%	42.13%
self-BLEU avg	48.50%	43.03%
self-ROUGE-1 A→B	64.83%	64.84%
self-ROUGE-1 B→A	72.63%	64.74%
self-ROUGE-2 A→B	47.99%	47.94%
self-ROUGE-2 B→A	58.69%	47.35%
self-ROUGE-L A→B	64.69%	64.71%
self-ROUGE-L B→A	72.46%	64.49%
style accuracy	43.02%	75.92%
acc-BLEU	45.76%	59.47%
g-acc-BLEU	45.68%	57.15%
h-acc-BLEU	45.60%	54.93%

TABLE V

Comparison of Base Sentences and Generated Style Transfers for the Neutral to Positive/Negative Amazon experiment.

<table>
<tr><th colspan="3">One to Many (Neutral → Positive, Negative)</th></tr>
<tr><th>Original (Neutral)</th><th>Generated (Positive)</th><th>Generated (Negative)</th></tr>
<tr><td>An average coffee at a decent price.</td><td>An excellent coffee at a reasonable price.</td><td>An awful coffee at a decent price.</td></tr>
<tr><td>Smaller than I realized.</td><td>THIS ITEM IS GREAT! FAST DELIVERY AND GREAT SELLER.</td><td>THESE CANS ARRIVED SO DENTED - I AM UNABLE TO OPEN THEM! COME ON!!!!</td></tr>
<tr><td>Not quite the taste I expected.</td><td>Not at all the best taste I expected.</td><td>Not at all the taste I expected.</td></tr>
<tr><th colspan="3">Many to One (Positive, Negative → Neutral)</th></tr>
<tr><th colspan="3">Positive → Neutral</th></tr>
<tr><td>Great tasting coffee.</td><td colspan="2">Not a fan of the flavor.</td></tr>
<tr><td>”Great candy, pricey!”</td><td colspan="2">”Good candy, mediocre.”</td></tr>
<tr><td>Delicious! We use it as high protein snack. Excellent product.</td><td colspan="2">Meh. We use it as high protein snack. Expensive.</td></tr>
<tr><th colspan="3">Negative → Neutral</th></tr>
<tr><td>I did not notice any difference in energy levels nor any weight loss. Disappointing for the price.</td><td colspan="2">I did not notice any difference in energy levels nor any weight loss. Overpriced for the price.</td></tr>
<tr><td>It is difficult to describe. Very unpleasant.</td><td colspan="2">It is difficult to describe. Quite chewy.</td></tr>
<tr><td>The flavor is very bad.</td><td colspan="2">The flavor is quite different.</td></tr>
</table>

TABLE VI

Validation metrics obtained over both configurations: A Neu|B Pos and A Neu|B Neg.

Metric	A Neu|B Pos	A Neu|B Neg
self-BLEU A→B	79.40%	79.62%
self-BLEU B→A	50.47%	69.94%
self-BLEU avg	64.94%	74.78%
self-ROUGE-1 A→B	89.61%	89.72%
self-ROUGE-1 B→A	71.81%	84.02%
self-ROUGE-2 A→B	83.29%	83.49%
self-ROUGE-2 B→A	57.89%	76.28%
self-ROUGE-L A→B	89.56%	89.68%
self-ROUGE-L B→A	71.71%	83.99%
style accuracy	37.94%	50.07%
acc-BLEU	51.44%	62.43%
g-acc-BLEU	49.64%	61.19%
h-acc-BLEU	47.90%	59.98%

TABLE VII
Performance metrics for DistilBERT Binary Classifier over Authorship style dataset.

Label	**Precision**	**Recall**	**F1-Score**	**Support**
trump	0.91	0.93	0.92	388
shakespeare	0.91	0.89	0.90	330
Overall Accuracy	-	-	**0.91**	**Tot**: 718

TABLE VIII
Comparison of performance metrics for both DeBERTa and DistilBERT Ternary Classifier over Authorship style dataset.

Label	**DeBERTa**			**DistilBERT**			**Support**
	Prec.	**Recall**	**F1-Score**	**Prec.**	**Recall**	**F1-Score**	
trump	86.00%	91.00%	88.00%	93.00%	86.00%	90.00%	388
lyrics	87.00%	85.00%	86.00%	77.00%	92.00%	84.00%	330
shakespeare	86.00%	82.00%	84.00%	92.00%	82.00%	87.00%	319
Accuracy	–	–	**86.00%**	–	–	**91.00%**	**Tot**: 1034

TABLE IX
Hyper-parameter settings for Shakespeare Binary and Ternary GAN experiments

Parameter	**Binary GAN**	**Ternary GAN**
Generator Model	BART-base	BART-base
Discriminator Model	DistilBERT-base-cased	DistilBERT-base-cased
Learning Rate	1e-4	5e-5
Batch Size	16	8
Number of Epochs	20	20
Max Seq Length	32	64

TABLE X
Comparison of plain and enhanced models' Generated sentences for A → B and B → A

Trump → Shakespeare		
Original	**Generated (Plain)**	**Generated (Enhanced)**
Democrats also want to bring in unlimited numbers of foreign refugees from overseas, all at the expense of supporting their own communities.	Nay, but first, let alone, all at the expense of their own lives.	Dost thou bring in yourself from thee, all at the expense of thine own?
The past administration didn't pay interest.	The pastime didn't pay homage.	The past gods didn't pay heed.
Some people on the political left are trying to impose their own views and ideas on everyone else, telling them what to believe and how to live.	Some villains on the street are wont to hang their own faces and eyes on everyone else, telling them what to believe and how to live.	Some on the earth bear their own sight And some on their foot, telling them what to hear.
Shakespeare → Trump		
Master, go on, and I will follow thee To the last gasp, with truth and loyalty.	They will go on, and we will follow them up to the last week, with those and thousands of people joining us.	President, go on, and I will follow closely to the last election, with truth and accuracy.
I am no villain.	I am no longer allowed to do this.	I am no politician.
He would not have spoke such a word.	He would not have received such a positive impact on the nation.	He wouldn't have said such a thing.

TABLE XI
METRICS COMPARISON OF BINARY TRAINING FOR NOT ENHANCED VS. ENHANCED GAN. THE HIGHEST VALUE FOR EACH METRIC IS BOLDED.

<table>
<tr><th rowspan="2">Metric</th><th colspan="2">A Tru |B Sha</th></tr>
<tr><th>Plain</th><th>Enhanced</th></tr>
<tr><td>self-BLEU A→B</td><td>**50.94%**</td><td>34.39%</td></tr>
<tr><td>self-BLEU B→A</td><td>**46.19%**</td><td>31.89%</td></tr>
<tr><td>self-BLEU avg</td><td>**48.57%**</td><td>33.14%</td></tr>
<tr><td>self-ROUGE-1 A→B</td><td>**72.80%**</td><td>58.99%</td></tr>
<tr><td>self-ROUGE-1 B→A</td><td>**69.47%**</td><td>56.80%</td></tr>
<tr><td>self-ROUGE-2 A→B</td><td>**60.34%**</td><td>42.66%</td></tr>
<tr><td>self-ROUGE-2 B→A</td><td>**55.33%**</td><td>38.82%</td></tr>
<tr><td>self-ROUGE-L A→B</td><td>**72.54%**</td><td>58.77%</td></tr>
<tr><td>self-ROUGE-L B→A</td><td>**69.34%**</td><td>56.56%</td></tr>
<tr><td>style accuracy</td><td>49.03%</td><td>**72.14%**</td></tr>
<tr><td>acc-BLEU</td><td>48.80%</td><td>**52.64%**</td></tr>
<tr><td>g-acc-BLEU</td><td>48.79%</td><td>**48.90%**</td></tr>
<tr><td>h-acc-BLEU</td><td>**48.79%**</td><td>45.42%</td></tr>
</table>

TABLE XII
COMPARISON OF BASE SENTENCES AND GENERATED STYLE TRANSFERS FOR THE LYRICS TO TRUMP/SHAKESPEARE EXPERIMENT.

<table>
<tr><th colspan="3">One to Many (Lyrics → Trump, Shakespeare)</th></tr>
<tr><th>Original (Lyrics)</th><th>Generated (Trump)</th><th>Generated (Shakespeare)</th></tr>
<tr><td>All the colors have faded.</td><td>All the colors have changed.</td><td>All the colors have jaded.</td></tr>
<tr><td>Cause sweetie, you’re my kinda guy.</td><td>Caesar, you’re my elder?</td><td>Caesar, you’re going to marry me.</td></tr>
<tr><td>I just wanna play a board game.</td><td>I just gotta play a board game.</td><td>I just gotta play a practice game.</td></tr>
<tr><th colspan="3">Many to One (Trump, Shakespeare → Lyrics)</th></tr>
<tr><th colspan="3">Trump → Lyrics</th></tr>
<tr><td>No, I’m only kidding.</td><td colspan="2">’Cause I’m only kidding.</td></tr>
<tr><td>They’ve taken thousands of killers and brought them back.</td><td colspan="2">They’ve took the outlaws and took them back.</td></tr>
<tr><td>Please refrain from doing so, Michigan.</td><td colspan="2">Prevent this beef from buzzin’.</td></tr>
<tr><td>They’re so nasty to me today.</td><td colspan="2">They’re so laid up on me today.</td></tr>
<tr><td>The entire American nation is standing by your side.</td><td colspan="2">The whole world is standing by your side.</td></tr>
<tr><td>Some of the results are incredible.</td><td colspan="2">Some of the songs are as long as the day.</td></tr>
<tr><td>And, despite all that we’ve all gone through.</td><td colspan="2">And dig through all that we’ve all gone through.</td></tr>
<tr><th colspan="3">Shakespeare → Lyrics</th></tr>
<tr><td>You see the hunger and oppression in your eyes.</td><td colspan="2">You see the hunger and the pain in your face.</td></tr>
<tr><td>I am no villain.</td><td colspan="2">I ain’t got no love.</td></tr>
<tr><td>Oh, my God, my dear, she’s the sweetest woman in the world.</td><td colspan="2">Oh, my darling, she’s the best.</td></tr>
<tr><td>Come, boy, with me.</td><td colspan="2">Come on, boy, with me.</td></tr>
<tr><td>No, I will not.</td><td colspan="2">No, I won’t be shaken.</td></tr>
<tr><td>What’s the matter with Romeo?</td><td colspan="2">What it’s like to be totally lost in love.</td></tr>
</table>

TABLE XIII
COMPARISON OF PERFORMANCE METRICS BETWEEN DISTILBERT AND DEBERTA-BASED TERNARY GAN MODELS FOR STYLE TRANSFER FROM LYRICS (A) TO TRUMP (B:TRU) AND SHAKESPEARE (B:SHA).

Metric	**DistilBERT Ternary GAN**		**DeBERTa Ternary GAN**	
	A:Lyr, B:Tru	**A:Lyr, B:Sha**	**A:Lyr, B:Tru**	**A:Lyr, B:Sha**
self-BLEU A→B	45.68%	45.47%	45.27%	44.96%
self-BLEU B→A	30.61%	26.75%	30.83%	27.58%
self-BLEU avg	38.14%	36.11%	38.05%	36.27%
self-ROUGE-1 A→B	69.24%	68.99%	69.39%	69.00%
self-ROUGE-1 B→A	54.22%	48.20%	53.73%	50.11%
self-ROUGE-2 A→B	57.75%	57.51%	57.56%	57.03%
self-ROUGE-2 B→A	42.09%	34.40%	41.78%	36.19%
self-ROUGE-L A→B	69.20%	68.92%	69.36%	68.96%
self-ROUGE-L B→A	53.95%	47.96%	53.58%	49.89%
style accuracy	52.68%	73.54%	48.59%	68.00%
acc-BLEU	45.41%	54.82%	43.32%	52.14%
g-acc-BLEU	44.83%	51.53%	43.00%	49.66%
h-acc-BLEU	44.25%	48.44%	42.68%	47.31%