Low-Resource Text Style Transfer for Bangla: Data & Models

Sourabrata Mukherjee[1], Akanksha Bansal[2], Pritha Majumdar[2]
Atul Kr. Ojha[3,2], Ondřej Dušek[1]
[1]Charles University, Faculty of Mathematics and Physics, Prague, Czechia
[2]Panlingua Language Processing LLP, India
[3]Insight SFI Centre for Data Analytics, DSI, University of Galway, Ireland
{mukherjee,odusek}@ufal.mff.cuni.cz
{akanksha.bansal,pritha.majumdar}@panlingua.co.in
atulkumar.ojha@insight-centre.org

Abstract

Text style transfer (TST) involves modifying the linguistic style of a given text while retaining its core content. This paper addresses the challenging task of text style transfer in the Bangla language, which is low-resourced in this area. We present a novel Bangla dataset that facilitates text sentiment transfer, a subtask of TST, enabling the transformation of positive sentiment sentences to negative and vice versa. To establish a high-quality base for further research, we refined and corrected an existing English dataset of 1,000 sentences for sentiment transfer based on Yelp reviews, and we introduce a new human-translated Bangla dataset that parallels its English counterpart. Furthermore, we offer multiple benchmark models that serve as a validation of the dataset and baseline for further research.

1 Introduction

Text style transfer (TST) aims to modify the style of a given text while preserving its underlying content (Shen et al., 2017; Prabhumoye et al., 2018; Li et al., 2018, see Figure 1). Prior research in text style transfer has primarily focused on the English language, overlooking languages with limited resources, such as Bangla. This work aims to close this gap specifically for Bangla and explores the text sentiment transfer task, which is a prominent subtask[1] of TST (Jin et al., 2022; Mukherjee et al., 2022; Luo et al., 2019a).

Bangla, also referred to as Bengali, is mostly spoken in the Indian regions of West Bengal, Assam, and Tripura and is the mother tongue of about 97.2 million speakers as per the 2011 Census Report of India.[2] It is one of the 22 scheduled (official) (Jha, 2010) Indian languages and the national language of Bangladesh. Syntactically, Bangla is agglutinative by nature. A single verb root in Bangla can have 150 + inflected forms (McCrae et al., 2021). There are multiple dialects of Bangla that vary mainly in terms of verb inflections and intonation (McCrae et al., 2021). For this work, we followed Bangla as spoken in West Bengal.

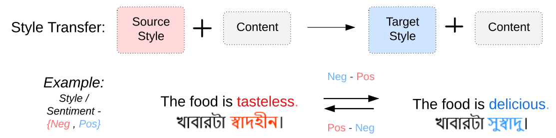


Figure 1: An example of sentiment transfer as a TST task in English and Bangla. Adapted from our previous paper (Mukherjee and Dusek, 2023).

The unique challenges posed by the low-resource nature of Bangla require specifically tailored innovative approaches. To achieve TST in Bangla, we build upon an existing English dataset of 1,000 sentences for this task adapted from Yelp reviews by Li et al. (2018). However, upon careful examination, we found that the quality of the original English dataset did not meet the standards we aimed to establish. To address this problem, we manually checked and modified the English dataset to improve its quality. Subsequently, we adapted the curated English dataset to the Bangla language, ensuring alignment in both content and structure. Importantly, we introduce a novel Bangla dataset, crafted by human annotators, serving as a parallel counterpart to the refined English dataset.

Furthermore, to facilitate the evaluation, we provide benchmark models capable of assessing the efficacy of text style transfer on our datasets. This paper marks a significant contribution to the field, as it not only pioneers text style transfer in the Bangla language but also provides a foundation for future research endeavors in multilingual text style transfer. Our work not only broadens the scope of text style transfer to include a low-resource language

[1]Moving forward, we will use the terms "style transfer" and "sentiment transfer" interchangeably in this paper.

[2]https://censusindia.gov.in/nada/index.php/catalog/42458

Proceedings of the First Workshop on Bangla Language Processing (BLP-2023), pages 34–47
December 7, 2023

but also underscores the importance of dataset quality. Our data and experimental code are released on GitHub.[3]

Our contributions are summarized as follows:

(i) We have enhanced the quality of the existing English parallel dataset for text sentiment transfer, improving its utility for research and applications.

(ii) We introduce a novel Bangla parallel dataset aligned with its English counterpart, effectively expanding the resources available for text style transfer in Bangla.

(iii) We present benchmark models to evaluate the performance of these datasets.

(iv) We also explored the challenging scenarios of having no style-parallel data or not using any human-annotated Bangla data for training (opting instead for English-to-Bangla machine translation). These experiments demonstrate the potential for meaningful results even with limited or no language-specific resources.

2 Related Work

Existing works in TST are mostly aimed at the English language and can be broadly classified into below categories:

TST with Parallel Data TST can be modeled as a sequence-to-sequence task and trained on pairs of texts with similar content but different styles. Here, Jhamtani et al. (2017) used a sequence-to-sequence model with a pointer network to translate modern English into Shakespearean English. Mukherjee and Dusek (2023) leveraged minimal parallel data and incorporated various low-resource methods to explore the TST task. However, this approach to TST is inherently challenging due to the scarcity of parallel data (Hu et al., 2022; Mukherjee and Dusek, 2023).

Non-Parallel Approaches to TST Two main strategies were employed to avoid reliance on parallel data: (i) straightforward text replacement, where style-specific phrases are explicitly identified and replaced (Li et al., 2018; Mukherjee et al., 2022), (ii) implicit style-content disentanglement via latent representations through techniques such as back-translation and autoencoding (Shen et al., 2017; Zhao et al., 2018; Fu et al., 2018; Prabhumoye et al., 2018; Hu et al., 2017). Adversarial learning was shown to improve the results of both approaches (Lample et al., 2019; Dai et al., 2019; Li et al., 2019; Luo et al., 2019b). Despite a lot of progress, non-parallel approaches tend to produce mixed results and often require large amounts of non-parallel data, which is not readily available for many styles, limiting their practical applicability in low-resource settings (Li et al., 2022).

Multilingual style transfer remains a relatively uncharted territory in prior research. In a comprehensive survey conducted by Briakou et al. (2021), only one work of TST was identified in languages such as Chinese, Russian, Latvian, Estonian, and French. Additionally, they introduced an evaluation dataset for formality transfer, encompassing French, Brazilian Portuguese, and Italian. Another study focused on formality transfer across various Indic languages (Krishna et al., 2022). Existing work has primarily concentrated on resource-rich languages, leaving languages like Bangla understudied in the domain of TST. The only previous work on Bangla known to us is the experiment of Palash et al. (2019), who used a small amount of non-parallel data to train an autoencoder, with largely negative results.

3 Dataset Creation

We utilized the Yelp dataset (Li et al., 2018), which is publicly available and has been used by prior TST experiments. It consists of user-generated content in the form of reviews for hospitality establishments. For each review sentence that is originally positive or negative, a parallel sentence has been created where the sentiment has been flipped but sentiment-independent content retained as much as possible. The dataset is in English. 500 sentences have been transferred from negative to positive and another 500 from positive to negative.

Implicit and Explicit Sentiment in Text Data The methodology behind creating sentences where sentiment transfer has taken place is a crucial process and a creative one. It primarily involves the identification of the sentiment-bearing attribute; for example, in the sentence *The food is tasteless*, "tasteless" is the sentiment-bearing attribute. The sentiment-bearing attribute can be transformed in

[3] Code: `https://github.com/souro/multilingual_tst`, data: `https://github.com/panlingua/multilingual-tst-datasets`.

multiple ways, here, e.g., by using an antonym of "tasteless" or adding a negation marker to tasty. The output, thus, could be either *The food is tasty* or *The food is not tasteless*. In both output sentences, the maximum lexical context has been preserved. It is the naturalness of the sentence or utterance that decides on preference between the two options. Sentiment-bearing attributes expressed as singular words or phrases are relatively easy to handle. The difficulty arises when the sentiment is carried or expressed implicitly. For this, the principle of sounding natural must be given prominence over lexical context preservation. A few such examples are reported in Table 3 in Appendix A.

English Data Correction The original English Yelp dataset included several discrepancies, some of which are reported in Table 4 in Appendix A: spelling mistakes, the incorrect sentiment of input sentences (flipped or neutral), compromise on naturalness, loss of context that could be preserved, or not changing the sentiment correctly in the target data, especially in cases where sentiment was expressed implicitly. For these reasons, we edited 451 sentences out of 1,000 in the original English Yelp dataset to meet the requirements of our experiment.

Creation of Bangla Data This dataset has been translated from English to Bangla to serve the aims of our experiment. Apart from usual translation challenges, specific problems arise for this particular dataset where sentiment transfer must be maintained (see Tables 5 and 6 in Appendix A for specific examples).

Some expressions that appear natural in English may come across as unnatural in Bangla. Hence, the complete lexical context may not be preserved during translation. Ambiguity in sentences poses difficulties in preserving multiple interpretations, and sentences with implied meanings are particularly challenging to translate. To address this, we often use similar phrases that maintain naturalness but may compromise lexical context. Slang words further complicate the translation process, as do instances where the original meaning is unclear, resulting in equally unclear translations. Consistency is crucial for small datasets, as variations in translation can affect results; for instance, bland can be translated as either 'flavourless' or 'tasteless'. Maintaining consistency was challenging in this respect. Lastly, a lack of cultural knowledge may lead to misinterpretations in translation, exemplified by cases like 'bs' meaning 'bullshit'.

4 Models

Our models are categorized into three approaches: parallel, non-parallel (not using parallel training data), and cross-lingual, i.e., without using our Bangla dataset for training. For an overview of the methodologies, see Figure 2.

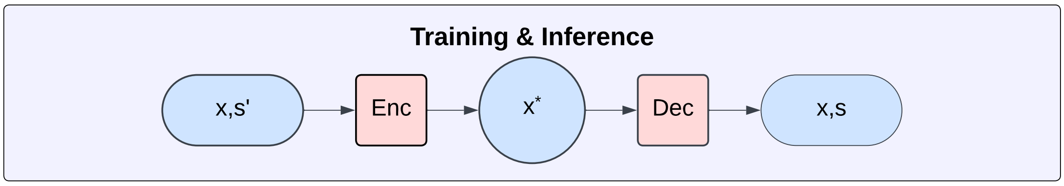
4.1 Parallel Style Transfer

Here, we simply fine-tune a pre-trained multilingual BART model (mBART) (Liu et al., 2020) using the parallel English and Bangla datasets constructed in Section 3. This approach is directly based on our previous work (Mukherjee and Dusek, 2023).

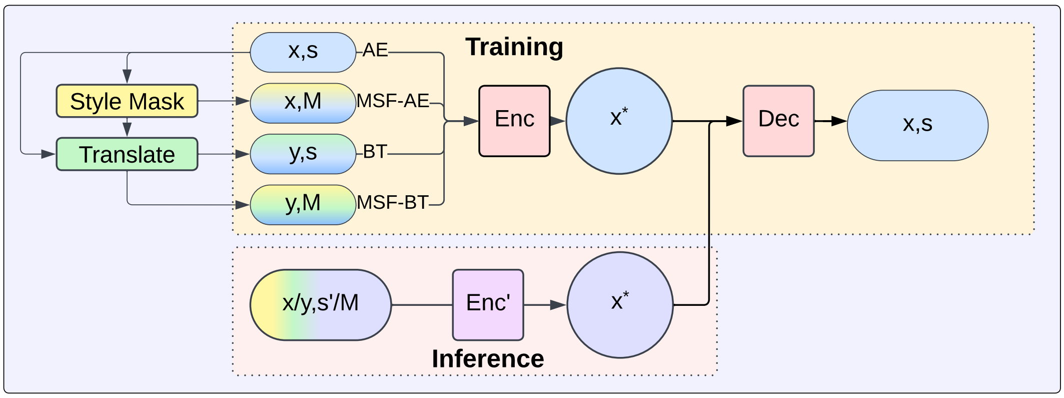
4.2 Non-parallel Style Transfer

In this experiment, we only use one part of the data at a time (positive/negative), never using the human-labelled targets for a given example. We harness the power of reconstruction of the input using an **auto-encoder** (AE) (Shen et al., 2017; Li et al., 2021) and **back-translation** (BT) (Prabhumoye et al., 2018; Mukherjee et al., 2022). In the BT process, for English sentences, we perform a cycle of translation, using English-to-Bangla-to-English, while for Bangla sentences, we apply Bangla-to-English-to-Bangla translation. For both AE and BT approaches, we train two separate models for each sentiment. At inference time, input is simply fed to the model trained on the intended target sentiment.

Masked Style Filling (MSF) We further extended the above AE and BT approaches by masking out the style-specific lexicon in the input sentence. Instead of relying on a fixed, contextually unaware style lexicon lookup, we take a dynamic and sentence-level perspective to identify important style-specific words in a sentence. This approach recognizes that words can have different stylistic roles based on the context in which they appear. To achieve this, we employ integrated gradients, a well-known model interpretability technique (Sundararajan et al., 2017; Janizek et al., 2021) on a fine-tuned mBERT (Pires et al., 2019) style classifier. This technique provides word attributions, essentially scores that show how much a word contributes to the style classifier model's prediction.



(1) Parallel Style Transfer



(2) Non-parallel Style Transfer

Figure 2: Overview of the Methodologies. (1) Parallel Sentences: This method employs aligned pairs of sentences with opposite styles, such as positive-to-negative and negative-to-positive. It employs a basic sequence-to-sequence (seq2seq) text generation approach, using an encoder (*Enc*) to process the input (x) and a decoder (*Dec*) to generate the opposite-style sentence (s'). For instance, to convert a positive sentence to a negative one, *Enc* encodes the positive text, and *Dec* decodes it into a negative sentiment. (2) Non-Parallel Data: In cases where aligned sentences are unavailable, this approach leverages non-parallel datasets containing positive and negative text. Two strategies are used: First, reconstruction, which uses auto-encoding (*AE*) or back-translation (*BT*). In *BT*, the input (x) is machine-translated to the opposite language (y) beforehand. Separate models are trained to reconstruct positive and negative sentences, but during inference, cross-models are used. For example, when transferring from positive to negative sentiment, the input is a positive sentence, and the model used for reconstruction is the one trained on negative sentences. The opposite applies to negative-to-positive transfers. In addition to this, Masked Style Filling (*MSF*) may be applied as preprocessing. *MSF* masks style-specific lexicon within the input, aided by a trained classifier and axiomatic attribution scores that identify style lexicon. The resulting style-masked sentence, denoted as (x, M) or (y, M), then undergoes the same reconstruction process (*AE* or *BT*).

With these word attribution scores in hand, we selectively mask out words that are considered style lexicon. We set a threshold to determine how much of the overall style should be removed from the sentence. The objective here is to create sentences that are "style-independent", devoid of specific stylistic markers. We then use these modified sentences as input to our AE and BT reconstruction models to reconstruct the original sentences. We again train two separate models for each sentiment and feed inputs to to the model trained on the intended target sentiment at inference time.

4.3 Cross-Lingual Style Transfer

We explore two basic cross-lingual alternatives that circumvent the use of the manually created Bangla dataset. Firstly, we employ English sentences from the parallel dataset, translate them into Bangla, and use this translated text for training. Secondly, we take the English output generated by the model trained on a parallel English dataset and translate it into Bangla. These cross-lingual approaches offer intriguing insights into multilingual text style transfer where the TST dataset is not available in the target language.

5 Experimetal Settings

Each dataset consists of 500 positive-to-negative and 500 negative-to-positive sentences (see Section 3). To maintain a consistent approach across all our experiments, we have divided these datasets into 400 examples for training, 100 for development, and 500 for testing purposes.

We used the mBART-large-50 model (Tang

et al., 2020) from the HuggingFace library (Wolf et al., 2020) for both English and Bangla. To optimize model performance, hyperparameter tuning was performed, resulting in the selection of a learning rate of 1e-5 and a batch size of 3. Dropout was applied with a rate of 0.1 across the network. Additionally, L2 regularization with a strength of 0.01 was introduced. Training ran over 5 epochs.

For machine translation in the BT reconstruction experiment, we use the Facebook NLLB-200-3.3B models (Costa-jussà et al., 2022) from HuggingFace.

For our MSF experiments and for evaluating sentiment transfer accuracy, we fine-tuned a classifier based on the BERT-base multilingual cased model (Devlin et al., 2018; Pires et al., 2019), using the same training set as our primary task. This fine-tuned classifier achieves accuracy rates of 87.0% for English and 83.0% for Bangla. In the MSF process, we employ a threshold of 0.5 to selectively filter style lexicon.

6 Evaluation and Results

6.1 Evaluation

The evaluation process encompasses three key aspects: sentiment transfer accuracy, content preservation, and fluency. To assess sentiment transfer accuracy, we used our finetuned mBERT classifier (see Section 5). Consistent with prior research (Jin et al., 2022; Hu et al., 2022; Mukherjee et al., 2023), content preservation is assessed through BLEU score (Papineni et al., 2002) and embedding similarity (Rahutomo et al., 2012) against the input sentences. The embedding similarity is determined using language-agnostic BERT sentence embedding (LaBSE) (Feng et al., 2022) in conjunction with cosine similarity. Evaluating fluency, especially for Bangla, poses a challenge as there are no good assessment tools available for indic languages (Krishna et al., 2022). Previous research has cautioned against using perplexity (PPL) for fluency, as it tends to favor unnatural sentences with common words (Pang, 2019; Mir et al., 2019). Despite these problems, we still include a basic fluency evaluation using perplexity (PPL) measured with a multilingual GPT model (Shliazhko et al., 2022).

6.2 Results

Automatic metric results are shown in the Table 1. Our scores are roughly in the same ballpark as our previous experiments on the Yelp data (Mukherjee et al., 2023) and with somewhat lower style accuracy but higher content preservation scores than most other previous works (cf. Li et al., 2020). However, a direct comparison on the English data is not possible due to our corrections of the dataset (see Section 3). Based on cursory manual checks of the output texts, the scores reflect the individual models' performance well.

Style Accuracy: The benchmark model utilizing the parallel dataset demonstrates strong style accuracy. However, in the case of Bangla, the accuracy drops in comparison to English, indicating potential challenges in Bangla-style transfer. Non-parallel data models, such as AE and BT, exhibit significantly lower style accuracy in both languages.

Content Preservation: While the parallel model and the MSF-AE model perform relatively well in both languages, other non-parallel models struggle to preserve content effectively. The MSF approach in general enhances content preservation, narrowing the gap slightly between parallel and non-parallel data models.

Comparison of AE and BT: When comparing the performance of AE and BT models, AE tends to outperform BT in content preservation, but BT outperforms AE in style transfer accuracy.

Impact of MSF: The introduction of the MSF approach in general improves the results of both AE and BT models, increasing style accuracy and fluency, but at the cost of content preservation.

Parallel vs. non-parallel Data: As expected, parallel data models consistently outperform their non-parallel counterparts across various metrics. However, the incorporation of the MSF approach mitigates some of the challenges posed by non-parallel data, highlighting its effectiveness in bridging the performance gap.

Cross-lingual Experiments: By not using the actual Bangla dataset entirely, we explored two alternative approaches: (i) translating parallel English training sentences to Bangla and (ii) translating the English style transfer output to Bangla. Interestingly, both methods yield competitive results in Bangla, showcasing the potential of the style-parallel English dataset and simple translation for the text style transfer task if the actual TST

	English				Bangla			
Models	**ACC**	**BLEU**	**CS**	**PPL**	**ACC**	**BLEU**	**CS**	**PPL**
Parallel Style Transfer								
Parallel	77.0	46.5	81.0	97.5	66.0	34.5	81.0	7.7
Non-parallel Style Transfer								
AE	13.0	42.0	78.0	102.2	17.0	31.0	78.0	7.8
BT	28.0	10.0	64.5	139.4	33.5	3.0	63.5	7.3
MSF-AE	59.5	37.5	75.5	136.0	72.0	26.5	72.5	7.9
MSF-BT	59.5	9.5	62.0	90.2	55.5	1.0	43.0	26.7
Cross-Lingual Style Transfer								
Train-En-TR		-			61.0	28.0	79.0	7.7
En-OP-TR		-			64.5	6.0	74.5	6.8

Table 1: Automatic evaluation results. We measure the sentiment classifier accuracy (ACC), BLEU score, Content Similarity (CS), and Fluency (PPL), see Section 6.1. We have several models (see Section 4): Parallel that uses parallel data, AE and BT for non-parallel data using the reconstruction approach, and the extended models MSF-AE and MSF-BT employing Masked Style Filling. Train-En-Tr involves training without the human-annotated Bangla dataset by using English-to-Bangla machine-translated training data. En-OP-TR refers to the Bangla translation of English output generated by mBART-base using parallel English data.

dataset is not available in Bangla for training purposes. The performance of these methods is on par with or surpasses that of the non-parallel Bangla dataset-based models, underscoring the viability of using machine translation in the pipeline.

Comparison of English and Bangla Results: While the scores in both languages are not directly comparable, overall lower values for Bangla show that this problem is likely more challenging here, not least due to Bangla's more complex morphology or lower amount of pretraining in the underlying mBART language model. Both languages however correlate relatively well in terms of the relative performance of the individual models (parallel model are the best, MSF improves scores, BT seems worse than AE on content preservation).

In conclusion, our experiments emphasize the significance of parallel data in text style transfer and highlight the benefits of the MSF approach. The choice of model depends on the specific language, task requirements, and availability. Generated output samples are shown in Table 2.

7 Conclusion

In this study, we delved into the challenging domain of text style transfer primarily for the Bangla language, addressing the scarcity of resources in it. This work contributes essential resources and benchmark models for both, Bangla and English. Future work involves exploring further underrepresented languages in the multi-lingual TST research.

Limitations

Data Bias: Our study relies on publicly available text data, which may inherently contain biases present in the sources from which it was collected. These biases can affect the performance of models trained on such data and may lead to biased outputs in sentiment transfer tasks.

Generalization: While our models demonstrate good performance on our datasets, their ability to generalize to other domains or contexts may be limited.

Subjectivity and Context: Sentiment analysis is inherently subjective, and the sentiment labels assigned to sentences may not universally apply. The context in which a sentence is used can significantly influence its sentiment, and our models may not always capture nuanced contextual variations.

Evaluation Metrics: While we have employed a variety of evaluation metrics, including style transfer accuracy, content preservation, and fluency, no single metric captures all aspects of sentiment transfer. The evaluation process remains an active area of research, and further advancements in metrics may be needed.

Ethics Statement

Data Privacy and Consent: We are committed to respecting data privacy and ensuring that all data

Models	**Negative → Positive**	**Positive → Negative**	**Analysis**
Reference	hate the aternoon-tea at the phoenician. → love the afternoon-tea at the phoenician. ফিনিশিয়ানে দুপুরের চা একদম অপছন্দের । → ফিনিশিয়ানে দুপুরের চা খুব পছন্দের ।	i love their fresh juices as well. → i don't like their fresh juices either. আমার তাদের তাজা ফলের রসও খুব পছন্দ । → আমার তাদের তাজা ফলের রসও একদম পছন্দ না ।	The examples have been chosen to exemplify the use of antonym and NEG-marker to flip the sentiments. In both the examples, the remaining lexical context remains preserved.
Parallel	love the aternoon-tea at the phoenician. ফিনিশিয়ানে দুপুরের চা খুব পছন্দের ।	i hate their fresh juices as well. আমার তাদের তাজা ফলের রসও খুব পছন্দ নয় ।	For both languages, the transfer is extremely smooth with only the sentiment-bearing attributes changed, and the lexical context preserved.
AE	hate the aternoon-tea at the phoenician. ফিনিশিয়ানে দুপুরের চা একদম অপছন্দের ।	i love their fresh juices as well. আমার তাদের তাজা ফলের রসও খুব পছন্দ ।	Basically the model was able to reconstruct the input successfully, thus preserving the content, for both Bangla and English, but it failed in transforming the sentiment fully.
BT	I like the morning coffee. সকালটাটাটা খুবই পছন্দের কাছ থেকে।।	I hate their cheese. How আমি তাদের সুস্বাদু পানীয় খেতে পছন্দ করি।	The sentiments in English have transformed as desired but the lexical context has been compromised. In Bangla, the sentiment transfer has failed and the context preservation is worse in Positive-Negative as compared to Negative-Positive. Note, that in English, the context transformation remains within their respective categories unlike Bangla.
MSF-AE	the aternoon - tea at the phoenician. দুর্দান্ত সেলফিফিচালত না ।	didn't love their fresh as well. আমার তাদের তাজা ফলের রসও খুব খারাপ ।	While the sentiment-bearing attribute gets dropped altogether in English Negative-Positive making it neutral, but the sentiments have successfully transformed in Positive to Negative. However, the lexical context is slightly compromised in the latter. In Bangla, no sentiment transfer took place in Negative-Positive while sentiments were successfully transferred from Positive to Negative. This model has performed better for Positive to Negative.
MSF-BT	I like the Mexican chicken. সকালের সূর্যসা খুবই অস্বাভাবিকভাবে পছন্দ করা হয়।	I hate them. আমি তাদের fresh খাবার পছন্দ পছন্দম ।	Much like BT above, the sentiments have transformed successfully but so has the lexical context. Therefore, context preservation in this model is erroneous.
Train-En-TR	ফিনিশিয়ানে দুপুরের খাবার খুব পছন্দের ।	আমার তাদের তাজা ফলের রস খুব পছন্দ নয়।	In Negative to Positive the task has been completed perfectly. In Positive to Negative, sentiment transfer has been successfully carried out but content preservation is average.
En-OP-TR	ফিনিকিয়ান এর সকালের চা ভালো লাগে।	আমি তাদের তাজা রস পছন্দ করি না।	The output is in accordance with the model's aim.

Table 2: Here are sample outputs from our models, with sentiment marker words highlighted. The outputs for both the positive-to-negative and negative-to-positive tasks align with the scores presented in Table 1. In both English and Bangla sentences for both tasks, the parallel model performs reasonably well in terms of sentiment transfer and content preservation. On the other hand, the non-parallel models, AE and BT, show below-average performance, but their MSF extensions attempt to enhance sentiment transfer accuracy. MSF-AE is decent in content preservation but struggles with sentiment preservation, while MSF-BT performs decent in sentiment transfer accuracy but falls short in content preservation. Overall, the generated English samples are slightly better than the Bangla ones.

used in our research is anonymized and devoid of personally identifiable information. We have taken measures to protect the privacy and confidentiality of individuals whose data may be included in our datasets.

Bias Mitigation: We acknowledge the potential presence of bias in our data sources and have taken steps to minimize the impact of such bias during model training and evaluation. We prioritize fairness and strive to mitigate any potential bias in our results.

Transparency and Reproducibility: We are dedicated to providing transparency in our research methods, including dataset collection, preprocessing, and model training. We encourage reproducibility by making our code and datasets publicly available.

Informed Consent: In cases where our research involves human annotators or data contributors, we have sought informed consent and have followed ethical guidelines for data collection and usage.

Social Impact: We recognize the potential social impact of our research and remain vigilant about the responsible use of AI technologies. We aim to contribute positively to the field of sentiment analysis and ensure our work benefits society as a whole.

By acknowledging these limitations and adhering to ethical principles, we aim to conduct responsible and impactful research in multilingual TST. We remain committed to addressing ethical concerns and improving the robustness and fairness of our models as we continue our research endeavors.

Acknowledgements

This research was supported by the European Research Council (Grant agreement No. 101039303 NG-NLG) and by Charles University projects GAUK 392221 and SVV 260575. We acknowledge of the use of resources provided by the LINDAT/CLARIAH-CZ Research Infrastructure (Czech Ministry of Education, Youth, and Sports project No. LM2018101). We would also like to acknowledge Panlingua Language Processing LLP for this collaborative research project and for providing the dataset.
Atul Kr. Ojha would like to acknowledge the support of the Science Foundation Ireland (SFI) as part of Grant Number SFI/12/RC/2289_P2 Insight_2, Insight SFI Centre for Data Analytics.

References

Eleftheria Briakou, Di Lu, Ke Zhang, and Joel Tetreault. 2021. Olá, bonjour, salve! XFORMAL: A benchmark for multilingual formality style transfer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3199–3216, Online. Association for Computational Linguistics.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *CoRR*, abs/2207.04672.

Ning Dai, Jianze Liang, Xipeng Qiu, and Xuanjing Huang. 2019. Style transformer: Unpaired text style transfer without disentangled latent representation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Volume 1: Long Papers*, pages 5997–6007, Florence, Italy.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 878–891. Association for Computational Linguistics.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 663–670.

Zhiqiang Hu, Roy Ka-Wei Lee, Charu C. Aggarwal, and Aston Zhang. 2022. Text style transfer: A review and experimental evaluation. *SIGKDD Explor.*, 24(1):14–45.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. 2017. Toward controlled generation of text. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1587–1596, Sydney, NSW, Australia.

Joseph D. Janizek, Pascal Sturmfels, and Su-In Lee. 2021. Explaining explanations: Axiomatic feature interactions for deep networks. *J. Mach. Learn. Res.*, 22:104:1–104:54.

Girish Nath Jha. 2010. The TDIL program and the Indian langauge corpora intitiative (ILCI). In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Eric Nyberg. 2017. Shakespearizing modern language using copy-enriched sequence to sequence models. In *Proceedings of the Workshop on Stylistic Variation*, pages 10–19, Copenhagen, Denmark.

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

Kalpesh Krishna, Deepak Nathani, Xavier Garcia, Bidisha Samanta, and Partha Talukdar. 2022. Few-shot controllable style transfer for low-resource multilingual settings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 7439–7468. Association for Computational Linguistics.

Guillaume Lample, Sandeep Subramanian, Eric Michael Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2019. Multiple-attribute text rewriting. In *7th International Conference on Learning Representations, ICLR 2019*, New Orleans, LA, USA, May 6-9, 2019.

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 2019, Hong Kong, China, November 3-7, 2019*, pages 3302–3311.

Jicheng Li, Yang Feng, and Jiao Ou. 2021. SE-DAE: style-enhanced denoising auto-encoder for unsupervised text style transfer. In *International Joint Conference on Neural Networks, IJCNN 2021, Shenzhen, China, July 18-22, 2021*, pages 1–8. IEEE.

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, NAACL-HLT 2018, Volume 1 (Long Papers)*, pages 1865–1874, New Orleans, Louisiana, USA.

Xiangyang Li, Xiang Long, Yu Xia, and Sujian Li. 2022. Low resource style transfer via domain adaptive meta learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022*, pages 3014–3026, s, WA, United States.

Xiao Li, Guanyi Chen, Chenghua Lin, and Ruizhe Li. 2020. DGST: a Dual-Generator Network for Text Style Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7131–7136, Online. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Trans. Assoc. Comput. Linguistics*, 8:726–742.

Fuli Luo, Peng Li, Pengcheng Yang, Jie Zhou, Yutong Tan, Baobao Chang, Zhifang Sui, and Xu Sun. 2019a. Towards fine-grained text sentiment transfer. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Volume 1: Long Papers*, pages 2013–2022, Florence, Italy.

Fuli Luo, Peng Li, Jie Zhou, Pengcheng Yang, Baobao Chang, Xu Sun, and Zhifang Sui. 2019b. A dual reinforcement learning framework for unsupervised text style transfer. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019*, pages 5116–5122, Macao.

John P McCrae, Atul Kr Ojha, Bharathi Raja Chakravarthi, Ian Kelly, Patricia Buffini, Grace Tang, Eric Paquin, and Manuel Locria. 2021. Enriching a terminology for under-resourced languages using knowledge graphs. *Electronic lexicography in the 21st century (eLex 2021) Post-editing lexicography*, page 67.

Remi Mir, Bjarke Felbo, Nick Obradovich, and Iyad Rahwan. 2019. Evaluating style transfer for text.

Sourabrata Mukherjee and Ondrej Dusek. 2023. Leveraging low-resource parallel data for text style transfer. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 388–395, Prague, Czechia. Association for Computational Linguistics.

Sourabrata Mukherjee, Vojtech Hudecek, and Ondrej Dusek. 2023. Polite chatbot: A text style transfer application. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: EACL 2023 - Student Research Workshop, Dubrovnik, Croatia, May 2-4, 2023*, pages 87–93.

Sourabrata Mukherjee, Zdenek Kasner, and Ondrej Dusek. 2022. Balancing the style-content trade-off in sentiment transfer using polarity-aware denoising. In *Text, Speech, and Dialogue - 25th International Conference, TSD 2022, Brno, Czech Republic, September 6-9, 2022, Proceedings*, volume 13502 of

Lecture Notes in Computer Science, pages 172–186. Springer.

Mehedi Hasan Palash, Partha Protim Das, and Summit Haque. 2019. Sentimental Style Transfer in Text with Multigenerative Variational Auto-Encoder. In *2019 International Conference on Bangla Speech and Language Processing (ICBSLP)*, pages 1–4.

Richard Yuanzhe Pang. 2019. Towards actual (not operational) textual style transfer auto-evaluation. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*.

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.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. 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, ACL 2018, Volume 1: Long Papers*, pages 866–876, Melbourne, Australia. Association for Computational Linguistics.

Faisal Rahutomo, Teruaki Kitasuka, and Masayoshi Aritsugi. 2012. Semantic cosine similarity. In *The 7th international student conference on advanced science and technology ICAST*, volume 4, page 1.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi S. Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6830–6841.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Vladislav Mikhailov, Anastasia Kozlova, and Tatiana Shavrina. 2022. mgpt: Few-shot learners go multilingual. *CoRR*, abs/2204.07580.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *CoRR*, abs/2008.00401.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi 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 M. 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, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 38–45. Association for Computational Linguistics.

Junbo Jake Zhao, Yoon Kim, Kelly Zhang, Alexander M. Rush, and Yann LeCun. 2018. Adversarially regularized autoencoders. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 5897–5906, Stockholm, Sweden.

A Appendix

ID	Input Sentence	Output Sentence	Analysis
1	guess she wasn't happy that we were asking the prices.	She was certainly happy to mention the prices.	Simply removing NEG-marker couldn't work as 'she was happy we were asking the prices' sounded unnatural.
2	my toddler found a dead mouse under one of the seats.	the place is clean and hygienic for kids and toddlers.	In this, sentiment is carried by the event of finding a dead mouse, hinting at the cleanliness and hygiene issues. Therefore, the context was removed and direct lexical markers - clean and hygienic - were introduced.
3	what were you going to charge me when i purchase a dozen bagels ?	surely you were going to charge me reasonably when i purchase a dozen bagels.	This is satiric in nature and expressed as a question with no expectation for an answer. Such sentences had to be flattened to transfer style.
4	it looks like blended up chicken put into the shape of grilled pieces.	the grilled chicken was made just perfect	The input sentence doesn't explicitly mentions what is good or bad. It simply related an event which is then perceived as negative.
5	what about the time i wasted?	service was quick and swift.	Again, it is not really a question but a comment on bad service veiled in the form of a question.
6	i should have stuck with sun chinese dining.	this was a very great place to dine in at.	Sarcasm is used to express displeasure, hence the entire lexical context was compromised during the transfer process.
7	was n't busy , no biggie .	was busy , no biggie .	Nothing here says if being busy was something good or bad.
8	there is a reason they can get you in fairly quickly.	This place is the most sought after.	Here the sentiment is implicit in the observation of the status of a venue, where the user uses sarcasm to mention that why you can get a table so quickly is that this place is not much preferred.

Table 3: Examples of handling implicit sentiments.

ID	Negative	Positive	Analysis
1	stopped by for soda after being at the hobby shop next door. After the hobby shop, I stopped in for a soda but hated it.	after the hobby shop i stopped in for a soda and enjoyed it.	The original sentence lacked sentiment, hence, a sentiment-bearing attribute had to be inserted.
2	i was so disgusted i could not ~~way~~ wait for the rest of the day.	i was so full i could not way for the rest of the day. I was so happy I could not wait for the rest of the day.	Spelling mistakes sometimes made it difficult to understand the sentence. In this particular example, the unclear context also increased difficulty.
3	i know i should have sent this back and walk out.	i know i shouldn't have sent this back and walked out. I relished my order.	Lexical context could not be preserved for the sake of naturalness.
4	i'm not one of the corn people .	i'm proud to be one of the corn people.	Cultural undertone resolution was a problem.
5	i got there, was seated pretty ~~quickly~~ late, and then couldn't chose my color.	we were seated quick as soon as we got there, then we glady chose colors.	The input was incorrectly identified as negative making both, the input and output positive, hence, one had to be manipulated for negative.
6	sadly, we've been to this long established restaurant many times.	the restaurant has been great throughout the years fortunately, we 've been to this long established restaurant many times.	The original input lacked sentiment.
7	liar, liar, pants on fire.	truth truth be told ! honest people	Proverbial expressions were difficult to deal with.
8	too bad it was at the expense of the other customers.	~~too bad~~ gladly, it wasn't at the expense of the other customers.	Here, the challenge was to stranger sentiment with as little loss of context as possible.
9	talk about false advertising so call before you go !	No need to call before you go. they are exactly what their advertising claims for them.	This is an example of a sentence where the sentiment is implicit and hence difficult to transfer.
10	so you aren't my problem.	don't worry, you weren't my problem. I'm glad you're not causing any concerns for me.	This is also an implicitly negative sentence, hence, difficult to transfer style as well as translate.
11	not sure, is that a good thing or a bad thing?	I bet it's a good thing, and not a bad thing.	Here an attempt was made to provide more clarity on the context.
12	when i first came to phx...yes this sounded indian to me. when i first came to phx...yes this sounded unpleasant to me.	when i first came to phx...yes this sounded american to me. when i first came to phx... yes this sounded pleasant to me.	When sentiment-bearing attributes were cultural signifiers, for example, here, 'indian' was made positive with 'american', we decided to work with pleasant and unpleasent.
13	you won't find a ~~better~~ worse selection in scottsdale.	you won't find a better selection in ~~arizona~~ scottsdale.	When the input sentence was incorrectly identified as negative, editing was required. It affected the decision-making process for the sentiment transfer.
14	if i could give zero stars i def would.	the stars was 5 plus If I could give more stars I def would.	We also made reasonable changes in the data where we did not want the model to establish a link between numbers carrying neg/pos relationship. For example, in the example below we didn't want zero-five relationship to be seen as a definite neg-pos relationship.

Table 4: Text Sentiment Transfer English dataset improvement challenges' Examples.

ID	Positive	Negative	Analysis
1	i highly recommend e & m painting. আমি অবশ্যই ই অ্যান্ড এমের পেইন্টিঙের সুপারিশ করব।	I highly recommend avoiding e & m painting. আমি একদমই ই অ্যান্ড এমের পেইন্টিঙের সুপারিশ করব না।	Here, instead of translating 'avoiding' NEG-marker was used to flip the sentiment to maintain naturalness.
2	everything is fresh and so delicious! সবকিছু খুব তাজা এবং সুস্বাদু ছিল !	everything is so stale and bland! সবকিছু খুব পুরনো এবং অসুস্বাদু ছিল।	Oft-repeated words like 'bland' in this example have been translated consistently. Compare the translation of 'bland' with number 7 in Table 6.
3	the variety of sushi rolls makes for a good eating. খাওয়ার জন্য বিভিন্ন ধরনের শুষি রোল রয়েছে।	There is limited variety for sushi rolls. শুষি রোলের জন্য সীমিত বৈচিত্র্য রয়েছে।	In this example the lexical context of 'for a good eating' has been dropped to maintain naturalness in Bangla translation.
4	thanks for making our special night an event to remember. আমাদের বিশেষ রাতটা এতো স্মরণীয় বানানোর জন্য অনেকধন্যবাদ।	thanks for making our special night so horrible. আমাদের বিশেষ রাতটা নষ্ট করে দেওয়ার জন্য অনেক ধন্যবাদ।	Retaining the word 'thanks' adds a sarcastic touch during the style transfer process. The same is a challenge to maintain in Bangla.
5	when i first came to phx...yes this sounded unpleasant to me. যখন আমি প্রথম পিএইচএক্সে এসেছিলাম… এটাআমার কাছে অপ্রীতিকর শুনিয়েছিল।	when i first came to phx... yes this sounded pleasant to me. যখন আমি প্রথম পিএইচএক্সে এসেছিলাম…এটা আমারকাছে প্রীতিকর শুনিয়েছিল।	The use of the word 'yes' sounds forced in Bangla, hence had to be avoided.
6	what the hell are you doing ? তুমি এটা কি খারাপ কাজ করছ ?	you're doing great তুমি এটা ভালো কাজ করছ।	The word 'hell' is a negative sentiment-bearing word that means other than the common noun hell.
7	but unfortunately the rude woman was the one checking us out. কিন্তু দুর্ভাগ্যবশত, অভদ্র মহিলাটিই আমাদের চেকআউট করছিলেন।	but fortunately the polite woman was the one checking us out. কিন্তু সৌভাগ্যবশত, ভদ্র মহিলাটিই আমাদের চেক আউটকরছিলেন।	The 'checking out' could mean checking out at the counter or a slang. The translation is force to dilute the ambiguity and maintain one meaning.
8	this place is a shit hole with shit service. এই জায়গাটা যেরকম বাজে সেরকমই বাজে এর পরিষেবা।	this place is very nice with great service. এই জায়গাটা যেমন ভালো তেমনি ভালো তার পরিষেবা।	Slang Words pose challenges in translation.

Table 5: English and Bangla Text Sentiment Transfer Examples (Positive to Negative).

ID	Negative	Positive	Analysis
1	but it probably sucks too ! কিন্তু এটাও সম্ভবত খুব খারাপ !	but it probably doesn't suck too ! কিন্তু এটাও সম্ভবত খুব একটা খারাপ নয় !	Here the negative sentiment bearing words is 'sucks' that does not have an exact translation in Bangla. Hence, an approximate word had to be used which limits the range of meaning 'sucks' carry in English.
2	Their chips are ok, but their salsa is really bland. তাদের চিপস ঠিক ছিল, তবে সালসাটা অসুস্বাদু ছিল।	Their chips are good and their salsa is really tasty. তাদের চিপস ভালো ছিল, এবং সালসাটা অসাধারণ ছিল।	Please refer to Example 4 in Table 5.
3	the wine was very average and the food was even less. ওয়াইনের স্বাদ মোটামোটি ছিল তবে খাবারের স্বাদ খুব খারাপ ছিল।	the wine was above average and the food was even better. ওয়াইনের স্বাদ ভালো ছিল তবে খাবারের স্বাদ আরও ভালো ছিল।	Here the subtlety has been compromised during the translation process. Both 'average' and 'even less' have been directly interpreted as 'bad'.
4	for the record i am not a good cook , i use seasoning ! মোট কথা আমি একজন ভালো রাঁধুনি নই, আমি শুধু মশলার সাহায্যে রান্না করি।	for the record i am a terrific cook, i use seasoning ! আসলে আমি একজন অসাধারণ রাঁধুনি, আমি সব মশলা দিয়ে রান্না করি।	Here Bangla translation uses 'spice' for 'seasoning' where seasoning is a broader term in English with no exact translation in Bangla.
5	this is an old worn out hotel. এটা একটা পুরনো, জীর্ণ হোটেল।	this is an old vintage hotel. এটা একটা পুরনো, ভিনটেজ হোটেল।	Although a Bangla translation for Vintage exists, yet transliteration was preferred not only to maintain consistency but also to retain the exact flavour of vintage and not deviate towards antiquity.
6	talk about false advertising so call before you go ! মিথ্যে বিজ্ঞাপনের কথা শুনছিলাম তাই যাওয়ার আগে ফোন করে নিও।	they are exactly what their advertising claims for them. তারা যেটা বিজ্ঞাপন করে ঠিক সেটাই।	To resolve implicit meaning in this sentence, similar phrases were used which preserved the naturalness but compromised on the lexical context.
7	not so much these days. আজকাল খুব একটা না।	much more these days. আজকাল আরো অনেক বেশি।	Here, the meaning is very unclear leading to an equally unclear translation
8	half of my head was over processed. আমার অর্ধেক মাথা আর কাজ করছিল না।	half of my head was processed well. আমার অর্ধেক মাথা এখনো কাজ করছিল।	To resolve implicit meaning in this sentence, similar phrases were used which preserved the naturalness but compromised on the lexical context.

Table 6: English and Bangla Text Sentiment Transfer Examples (Negative to Positive).