A Comparative Analysis of Activation-Based Steering and Prompt-Based Editing for Text Style Transfer

Sonja Gievska
Martina Toshevska
Nikola Alchev 211051
Faculty of Computer Science
& Engineering, UKIM

Abstract—This paper presents a comparative analysis of two text style transfer approaches: Activation-Based Style Vector Steering and Prompt-Based Editing. The activation-based approach modifies the internal state of the model by injecting style vectors derived from differences in hidden activations of positive and negative Yelp reviews. On the other hand, the prompt-based approach applies iterative text editing guided by a masked language model and hill-climbing search. Both methods were evaluated on the Yelp sentiment dataset, with performance measured in terms of sentiment shift, grammatical correctness, and semantic similarity. The results show that activation-based steering achieves strong sentiment transfer with moderately fluent outputs, while prompt-based editing preserves the original meaning more reliably, though it leads to weaker sentiment changes. These findings highlight the trade-offs between internal model steering and external sentence editing in controllable text generation.

I. Introduction

A. Problem and Motivation

Changing the style of text while keeping its meaning is an important task in natural language processing. One common example is sentiment transfer, where the goal is to take text that sounds negative and make it sound positive, or the other way around. This problem is not easy, because the model needs to change the style of the text without making the sentence ungrammatical or changing the meaning too much.

Researchers have tried different ways to solve this. One way is to change the internal states of a language model so that it produces sentences with the desired style. Another way is to use prompt-based editing, where the model takes an existing sentence and makes small changes to rewrite it into the new style. Even though both of these methods work, it is still not clear which one is better for sentiment transfer, and what the trade-offs are.

B. Goal and Objectives

The main goal of this paper is to compare activation-based steering and prompt-based editing for sentiment transfer. In the activation-based method, we use GPT-Neo 1.3B and change its hidden states so that, when prompted, it generates sentences with positive or negative sentiment. In the prompt-based method, we use GPT-Neo 2.7B together with a masked language model to rewrite existing sentences step by step until the sentiment is changed.
The objectives are:

1. To test how well each approach changes the sentiment of text.
2. To measure how fluent and grammatical the new sentences are.
3. To check how well the meaning is preserved.

By doing this, we want to show the strengths and weaknesses of each method and see which situations they are best suited for.

C. Relation to Existing Research

This paper combines ideas from two active research areas. The first, the activation-based approach, is based on the idea that models store meanings and styles in certain directions inside their internal activations [1]. The second, the prompt-based editing approach, treats style transfer as a guided search problem using prompts to classify style [2]. In this report, we compare these two different approaches directly on the same task and dataset.

II. Related Work

One of the most recent approaches to controllable text generation is described in the paper Style Vectors for Steering Generative Large Language Models [1]. This work introduces the activation-based steering method. The main idea is to compute style vectors from the hidden activations of a transformer model. First, a large set of positive and negative sentences from the Yelp dataset is passed through the model. Then, the mean activations for positive sentences and for negative sentences are calculated, and their difference forms a style vector. During text generation, this style vector is injected into certain transformer layers, which pushes the model to generate sentences with the desired sentiment. Importantly, this method does not edit an existing input sentence. Instead, it influences the generation process so that, given a prompt, the model naturally produces outputs that follow the target style.

A different approach is offered in the paper Prompt-Based Editing for Text Style Transfer [2]. Here the problem is treated as a sentence rewriting task. The method uses a masked language model, such as RoBERTa, to propose edits to an input sentence. These edits can be insertions, deletions, or word replacements. Each candidate sentence is then scored based on three factors: how well it matches the target sentiment, how fluent and grammatical it is, and how similar it remains to the original sentence. A hill-climbing algorithm is applied to gradually improve the sentence until it reaches the desired sentiment. This approach, also evaluated on the Yelp dataset, focuses on preserving the meaning of the original text while shifting its style.

Both approaches make use of the Yelp dataset, which contains a large number of positive and negative review sentences. This dataset is especially suitable for sentiment style transfer because it provides clear examples of contrasting styles. In the activation-based method, Yelp reviews are used to extract the positive and negative activations that form the style vector. In the prompt-based editing method, the same dataset provides original sentences that can be rewritten into the opposite sentiment. Using this dataset ensures that both methods can be fairly compared under the same conditions for the task of sentiment transfer.

III. Experimental Setup

The main goal of this project is to compare two different approaches for sentiment style transfer: activation-based steering and prompt-based editing. The comparison is based on three evaluation measures: sentiment change, fluency and grammatical correctness, and semantic similarity.[1]

A. *Activation-Based approach*

In the activation-based approach, experiments were performed with GPT-Neo 1.3B. A large sample of 50,000 positive and 50,000 negative reviews from the Yelp dataset was used to extract hidden activations from the model. The mean activations for the positive and negative examples were calculated, and their difference formed the style vector. During text generation, this style vector was injected into the transformer, specifically into layers 12 to 18. Following the findings in Style Vectors for Steering Generative Large Language Models [1], the middle layers of transformer models tend to capture the strongest stylistic and semantic information, making them the most effective for steering. A steering scale of 0.3 was applied, providing a moderate but noticeable influence on the generation process.

The performance of this method was evaluated with three measures. Sentiment strength was measured using a RoBERTa-based sentiment classifier (VictorSanh/roberta-base-finetuned-yelp-polarity). Perplexity, which serves as a proxy for grammatical correctness and fluency, was computed with GPT-Neo 1.3B itself. Finally, semantic similarity was measured using the SentenceTransformers model all-MiniLM-L6-v2. For this method, semantic similarity was interpreted as the relevance of the generated answers to the input prompts, since the activation-based approach does not rewrite existing sentences but instead produces entirely new generations under sentiment steering.

B. *Prompt-Based Editing approach*

The prompt-based approach followed the method described in the paper Prompt-Based Editing for Text Style Transfer [2]. Here, GPT-Neo 2.7B was combined with a RoBERTa model as an editor. The input sentences came from the Yelp test set, which were then rewritten in both directions, meaning sentences were converted from positive to negative and from negative to positive. The editor model was *roberta-large*, which proposed candidate edits through masked token prediction. The editing process allowed three types of operations: word replacement, insertion, and deletion. A hill-climbing search was used to select the best candidate sentences, with a maximum of 6 steps per sentence and early stopping enabled if the target sentiment was reached earlier. The scoring function for the search was weighted across three dimensions: style weight 12, fluency weight 4, and semantic similarity weight 4. Evaluation was carried out using the same three measures as the activation-based approach: sentiment classification with Roberta, perplexity with GPT-Neo 1.3B, and semantic similarity with *all-MiniLM-L6-v2*. Unlike the activation-based method, semantic similarity here measures how close the rewritten sentence is to the original input, since prompt-based editing always starts from a given sentence.

IV. Results and Discussion

A. *Evaluation Metrics*

The experiments were evaluated along three dimensions:

1. Sentiment accuracy, measured with a RoBERTa-based classifier, which checks if the output sentence matches the target sentiment.

2. Perplexity, measured with GPT-Neo 1.3B, which serves as a proxy for grammatical correctness and fluency.

3. Semantic similarity, measured with all-MiniLM-L6-v2. For the activation-based method, similarity shows how relevant the generated output is to the input prompt. For the prompt-based editing method, similarity shows how close the rewritten sentence is to the original one.

[1] *Code link https://github.com/NikolaAlchev/text-style-transfer-steering*

B. *Activation-Based Steering Results*

TABLE I.

Metrics	Sentence Type	
	Factual Sentences	***Subjective Sentences***
Baseline Sentiment	0.72	0.74
Steered Sentiment	**0.89**	**0.92**
Baseline Perplexity	6.11	7.09
Steered Perplexity	6.59	8.53
Baseline Semantic Similarity	0.59	0.53
Steered Semantic Similarity	0.56	0.49

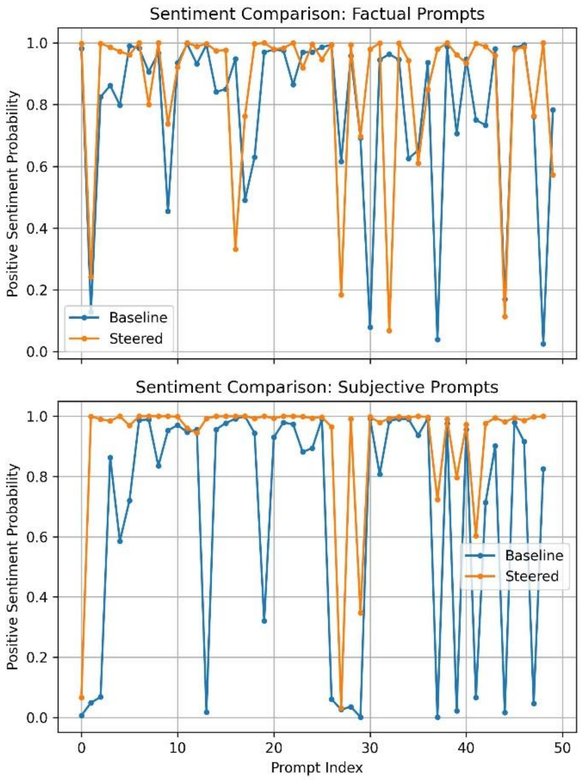

Figure 1. Sentiment comparison for Factual vs. Subjective prompts

The activation-based method shows a strong improvement in sentiment control for both factual and subjective prompts. Sentiment scores increased significantly after steering, indicating that the style vectors effectively pushed the model toward the desired polarity. Based on the results, we can see that subjective prompts exhibit a more noticeable sentiment shift compared to factual prompts. The increase in perplexity is relatively minor, suggesting that fluency was largely preserved. Semantic similarity decreased slightly, showing some drift from the original prompt. If the steering scale were increased further, even higher sentiment scores could be achieved, but this would likely come at the cost of higher perplexity and reduced semantic similarity.

C. *Prompt-Based Editing Results*

TABLE II.

Metrics	Sentiment Direction	
	Negative → Positive	***Positive → Negative***
Baseline Sentiment	0.08	0.97
Steered Sentiment	**0.56**	**0.51**
Baseline Perplexity	394.93	354.18
Steered Perplexity	159.01	449.133
Semantic Similarity	0 .64	0.83

The prompt-based editing method achieved only moderate improvement in sentiment transfer. For negative → positive, sentiment scores increased from 0.08 to 0.56, while for positive → negative they dropped from 0.97 to 0.51, which is still weaker than the activation-based approach. Fluency was inconsistent: perplexity decreased to 159.01 for negative → positive but rose to 449.13 for positive → negative. This is partly because the baseline sentences already had high perplexity, making grammatical correctness harder to maintain. On the other hand, semantic similarity stayed relatively high, meaning the rewritten sentences mostly preserved the original content.

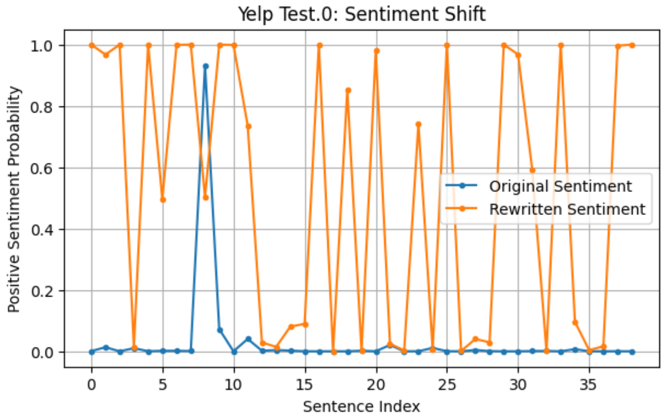

Figure 2. Sentiment shift for prompt-based editing: distribution of sentence sentiment from negative to positive.

V. Conclusion

This project compared two approaches for sentiment style transfer: activation-based steering with GPT-Neo 1.3B and prompt-based editing with GPT-Neo 2.7B, using the Yelp dataset and evaluating sentiment, fluency, and semantic similarity.

The results show clear differences between the two methods. Activation-based steering generally provides stronger control over sentiment, successfully shifting many sentences to the target sentiment. The strength of this effect varies depending on the prompt, with subjective prompts showing a more noticeable sentiment shift than factual prompts. Prompt-based editing, on the other hand, preserves the original sentence more reliably, but its ability to change sentiment is inconsistent, some outputs achieve the target sentiment, while others remain largely unchanged.

Overall, activation-based steering is more effective for controlling sentiment, while prompt-based editing is safer when preserving meaning is the priority. These findings highlight the trade-offs between intervening inside the model versus editing externally. It is also worth noting that the original studies used stronger GPT-J 6B models, whereas this project used smaller GPT-Neo models. This may explain some of the variation and less consistent results, suggesting that more powerful models could further improve performance for both approaches.

REFERENCES

[1] K. Konen, S. Jentzsch, D. Diallo, P. Schütt, O. Bensch, R. El Baff, et al., “Style vectors for steering generative large language models,” in Findings of the Association for Computational Linguistics: EACL 2024, St. Julian’s, Malta, Mar. 2024, pp. 782–802.

[2] G. Luo, “Prompt-based editing for text style transfer,” in Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, Dec. 2023, pp. 5740–5750.