CTRL + Style: An Analysis of Literary Representation and Style Transfer in Transformer Models

Alex Tulip

Contributing authors: alex.tulip@studenti.unimi.it;

Abstract

Keywords: Natural Language Processing, Text Style Transfer, Literary Analysis, Transformer Models, Computational Stylometry

1 Introduction

The ability to understand and replicate nuanced human expression is a key challenge for large language models (LLMs). Literary style, in particular, represents a complex combination of vocabulary, syntax, tone, and narrative structure that is easy for humans to recognize but difficult for machines to quantify. This project explores the intersection of natural language processing and literary analysis by investigating how modern transformer models internally represent the distinct authorial voices of Fyodor Dostoevsky and Charles Dickens. By examining the vector space representations of their works and attempting controlled style transfer, we aim to shed light on both the capabilities and the inherent biases of these powerful models when confronted with the subtleties of literary art.

2 Research Question and Methodology

2.1 Research Question

This study aims to answer the following research question: **"How do transformer models internally represent the stylistic differences between Dostoevsky's psychological realism and Charles Dickens's descriptive social commentary,**

and can these representations be manipulated for controlled text style transfer without altering the original text's semantic content?"

2.2 Measurable Objectives

Our research is guided by two primary objectives:

1. **Demonstrate and Quantify Stylistic Separation:** To prove that a transformer model's internal representations can effectively distinguish between the literary styles of Dostoevsky and Dickens. Our target is to achieve a classification accuracy of over 85% based on these representations.
2. **Achieve Controlled Style Transfer:** To successfully alter the style of a given text while preserving its core semantic content, evaluated using both automated metrics and qualitative analysis.

2.3 Methodology

Our methodology is divided into three main phases: data curation, text embedding, and stylistic analysis.

2.3.1 Dataset and Preprocessing

A robust and clean dataset is fundamental to this analysis. We constructed a custom corpus from the full texts of major works by our chosen authors, sourced from the Project Gutenberg archive.

- **Corpus Composition:**
  - **Fyodor Dostoevsky:** *Crime and Punishment*, *The Brothers Karamazov.*
  - **Charles Dickens:** *A Tale of Two Cities*, *Great Expectations*, *Oliver Twist.*
- **Preprocessing Pipeline:** A rigorous, multi-step cleaning process was applied to isolate the author's original writing. This involved programmatically removing Project Gutenberg headers and footers, trimming translator's notes and prefaces, and excising tables of contents. The cleaned text for each book was then segmented into paragraph-level chunks and normalized to create the final analysis-ready corpus.

2.3.2 Text Embedding with RoBERTa

To convert the textual data into a machine-understandable format, we employed a pre-trained RoBERTa-base model. Each paragraph-chunk from our corpus was fed into the model to generate a 768-dimensional vector embedding. We used the embedding of the `[CLS]` token as the representative vector for each chunk, a standard practice for capturing sentence- and paragraph-level semantics.

2.3.3 Stylistic Analysis

To test our primary hypothesis—that the embeddings of the two authors are separable—we performed a three-part analysis:

1. **t-SNE Visualization:** We used t-Distributed Stochastic Neighbor Embedding (t-SNE) to visualize the embeddings in a 2D scatter plot.
2. **Quantitative Classification:** We trained a Logistic Regression classifier on the embeddings to quantify the separability.
3. **Conceptual Interpretation:** We used Concept Activation Vectors (CAV) to investigate if the model's separation of styles corresponded to human-understandable literary concepts.

3 Experimental Results

Our experiments are divided into two main parts. First, we analyze the separability of the authors' styles in a model's embedding space. Second, we test the ability of a large language model to perform controlled style transfer.

3.1 Analysis of Stylistic Representation

3.1.1 Visualizing Stylistic Separation

The t-SNE visualization of the paragraph embeddings provides strong visual evidence of stylistic separation. As shown in Figure 1, the embeddings form two largely distinct clusters, with minimal overlap between them. This visual separation suggests that the RoBERTa model has successfully captured distinct stylistic features for each author.

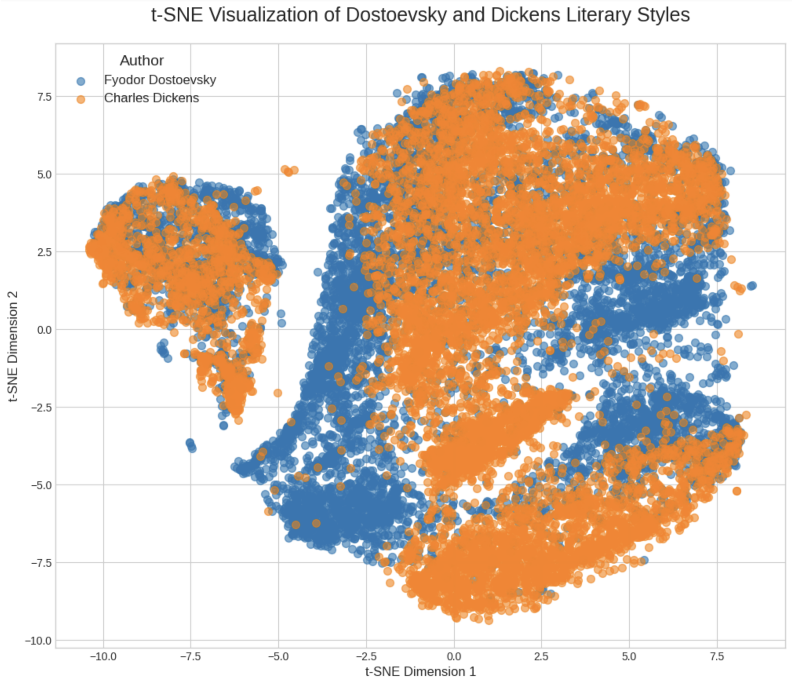

Fig. 1 t-SNE visualization of paragraph embeddings from Dostoevsky (blue) and Dickens (orange).

3.1.2 Quantifying Stylistic Separation

To move beyond visual intuition, we quantified the separability of the styles using a classification task. A Logistic Regression model was trained to predict the author from a paragraph's embedding alone. The model achieved an overall **accuracy of 92%** on the held-out test set, significantly surpassing our target objective of 85%. The detailed performance metrics in Table 1 and the confusion matrix in Figure 2 provide strong quantitative evidence that the model has learned a meaningful and separable internal representation of literary style.

Table 1 Classification report for predicting the author from text embeddings.

	precision	recall	f1-score	support
Fyodor Dostoevsky	0.93	0.91	0.92	1604
Charles Dickens	0.92	0.94	0.93	1816
accuracy			0.92	3420
macro avg	0.92	0.92	0.92	3420
weighted avg	0.92	0.92	0.92	3420

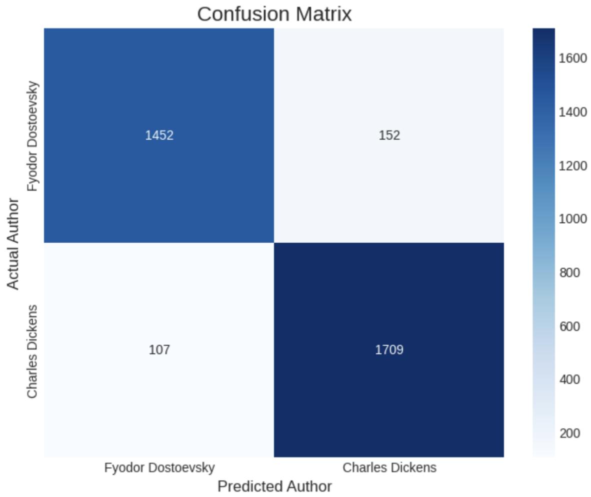

Fig. 2 Confusion matrix for the author classification task.

3.1.3 Interpreting Stylistic Dimensions with CAVs

Using the CAV technique, we tested the alignment of each author's work with two concepts: "Psychological Focus" and "Social Commentary." The results (Figure 3) reveal a clear conceptual divergence. Dostoevsky's texts showed a significantly higher sensitivity to the ”Psychological Focus” concept, while Dickens’s texts showed a much stronger alignment with ”Social Commentary,” indicating that the model’s internal representations are linked to the core thematic elements of each author.

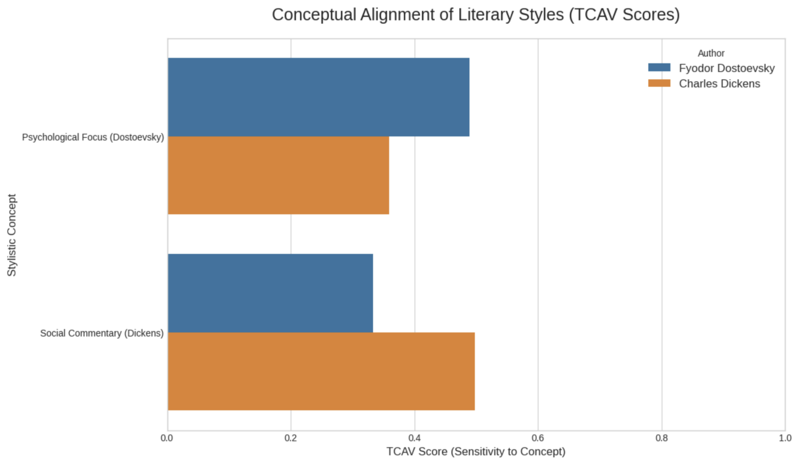

Fig. 3 Conceptual Alignment of Literary Styles (TCAV Scores).

3.2 Controlled Text Style Transfer

To address the second part of our research question, we explored the use of a large language model (`gemini-2.5-flash-lite-preview-06-17`) for prompt-based style transfer. We designed a detailed prompt with stylistic rules, negative constraints, and examples to guide the model. The performance was evaluated quantitatively across three metrics: Style Accuracy, Content Preservation (self-BLEU), and Fluency (Perplexity).

The results, shown in Table 2, reveal a significant asymmetry in the model’s ability to perform style transfer between the two authors.

Table 2 Quantitative Evaluation of Style Transfer Performance (Averages).

From	To	Style Accuracy	self-BLEU	Perplexity
Fyodor Dostoevsky	Charles Dickens	0.84	0.069	45.72
Charles Dickens	Fyodor Dostoevsky	0.20	0.066	70.79

The transfer from Dostoevsky’s minimalist style to Dickens’s ornate style was highly successful, achieving a **Style Accuracy of 84%**. This indicates that the model is very capable of the *additive* task of enriching text with descriptive language. Conversely, the transfer from Dickens to Dostoevsky proved to be a significant challenge.

The model achieved a **Style Accuracy of only 20%**, demonstrating a clear failure to adopt the target style. This suggests that the *constrictive* task of enforcing a minimalist, psychologically tense prose is much more difficult for the LLM.

4 Concluding Remarks

This project set out to investigate whether the abstract concept of literary style could be represented and manipulated by modern transformer models. Our experiments have yielded clear and insightful results.

The first phase of our research successfully demonstrated that a pre-trained RoBERTa model can indeed learn a meaningful, internal representation of literary style, achieving a classification accuracy of 92%. Our analysis with CAVs suggests that this separation is not arbitrary but is linked to the core thematic elements that define each author.

The second phase revealed a significant asymmetry in style transfer performance. The model was highly effective at the *additive* task of transferring a minimalist style to an ornate one (84% accuracy) but struggled significantly with the *constrictive* task of adopting Dostoevsky's tense, minimalist prose (20% accuracy).

This asymmetry is the most significant finding of our work. It suggests that while LLMs are adept at adding stylistic complexity, they have a fundamental difficulty in shedding their inherent verbosity to replicate sophisticated minimalism. This highlights a key challenge and a limitation in the current state of controlled text style transfer.

4.1 Future Work

Our findings open up several avenues for future research. One could explore more advanced style transfer techniques beyond simple prompting to see if they can better handle constrictive stylistic tasks. Another interesting direction would be to apply this methodology to a wider range of authors or even other languages to see if this additive/constrictive asymmetry is a universal phenomenon. Finally, a deeper dive into the specific linguistic features (e.g., sentence length distribution) that the model fails to replicate could provide even more granular insights into its limitations.

Appendix A AI Usage Disclaimer

Parts of this project have been developed with the assistance of Google's Gemini. The AI was used to support the structuring of methodological workflows, the drafting of descriptive texts, the generation of Python code for analysis, and the refinement of experimental conclusions. All content produced with AI assistance has been carefully reviewed, edited, and validated by me. I take full responsibility for the final content and its accuracy, relevance, and academic integrity.