CTRL+Style: Literary Style Analysis and Transfer between Jane Austen and Mark Twain

Daniyar Abdrakhmanov

Department of Computer Science, Università degli Studi di Milano, 20133, Milan, Italy.

Author's E-mail: daniyar.abdrakhmanov@studenti.unimi.it;

Abstract

This project examines how literary style can be represented and transformed using modern NLP methods. We focus on two contrasting authors, Jane Austen and Mark Twain, and analyse whether their stylistic differences appear in sentence-embedding spaces, classical stylometric features, and TF–IDF lexical markers.

Our pipeline includes passage segmentation, stylometric extraction, embedding projection with PCA and UMAP, and supervised style classification. We also perform bidirectional style transfer using a pretrained large language model and evaluate the outputs through a style-likeness classifier, cosine similarity, and readability shifts.

Results show clear stylistic separation between the two authors and demonstrate that LLMs can reproduce several target-style characteristics while largely preserving semantic content, though fine-grained control remains limited.

Keywords: Text Style Transfer, Literary Style Analysis, Stylometry, Sentence Embeddings, Natural Language Processing

1 Introduction

Understanding how literary style is represented and can be manipulated is a long-standing topic in both computational linguistics and literary studies. Style encompasses lexical choices, syntactic patterns, rhythm, and narrative structure, and its analysis has traditionally relied on stylometric indicators such as sentence length, vocabulary richness, or part-of-speech distributions. With the advent of transformer-based language models, it has become possible to study style through high-dimensional embedding spaces and to explore controllable style transfer using prompt-based generation.

In this project we investigate the stylistic contrast between two highly distinctive English-language authors, Jane Austen and Mark Twain. Our goal is to examine whether their writing styles form separable patterns across three complementary perspectives: classical stylometry, surface-level lexical markers, and sentence-level transformer embeddings. We further test whether a large language model can rewrite passages in the style of the other author while preserving semantic content.

The study was conducted within the NLP course of the Master's Degree in Computer Science at the University of Milan and aims to provide a clear, data-driven analysis of literary style as well as a practical evaluation of modern style transfer methods.

2 Methods

In this section we describe the data, feature representations, models, and evaluation protocol used in our study.

2.1 Corpus and Passage Segmentation

We constructed a parallel literary corpus from six English novels: three by Jane Austen (”Pride and Prejudice”, ”Emma”, ”Sense and Sensibility”) and three by Mark Twain (”Adventures of Huckleberry Finn”, ”The Prince and the Pauper”, ”A Connecticut Yankee in King Arthur's Court”). All texts were downloaded from Project Gutenberg and cleaned by removing boilerplate metadata, licensing headers, and trailing material.

Each novel was segmented into passages of approximately 150–300 words using spaCy sentence segmentation. Passages that were too short or too long were discarded. The resulting dataset contains a balanced number of passages per author. We then created a stratified split of 80% for training, 10% for validation, and 10% for testing, and saved all subsets as CSV files to ensure reproducibility.

2.2 Stylometric Feature Extraction

For each passage we computed a set of interpretable stylometric features. These include basic length measures (token count, sentence count, average sentence length, average word length), lexical richness (type–token ratio), part-of-speech ratios (nouns, verbs, adjectives, adverbs, proper nouns, pronouns), punctuation frequencies (exclamation and question marks per 1000 tokens), and the Flesch–Kincaid readability grade. Features were extracted with spaCy and the `textstat` library and aggregated into a feature matrix used both for exploratory analysis and for supervised classification.

2.3 Sentence Embeddings and Dimensionality Reduction

To obtain dense semantic and stylistic representations, we encoded each passage with the `all-MiniLM-L6-v2` SentenceTransformer model, resulting in 384-dimensional embeddings. Embeddings were L2-normalised. For qualitative analysis we applied

Principal Component Analysis (PCA) and Uniform Manifold Approximation and Projection (UMAP) on the training and validation embeddings.

2.4 Style Classification and Lexical Markers

We trained two types of style classifiers. First, a Linear Support Vector Machine (SVM) was fitted on stylometric features. Second, another Linear SVM was trained on the sentence embeddings. The latter classifier was used as a style-likeness oracle for evaluating generated passages.

In addition, we applied TF–IDF vectorisation over uni- and bi-grams with sublinear term frequencies, English stopword removal, and document frequency thresholds. A logistic regression classifier trained on TF–IDF features provided interpretable surface-level markers of style.

2.5 Style Transfer Generation and Evaluation

We generated stylistic rewrites using a pretrained Flan-T5 model in a purely prompt-based setup. For each direction (Austen→Twain and Twain→Austen), we sampled test passages and provided author-specific style guidelines.

Generated text was evaluated using three metrics: (1) a style score from the embedding-based SVM classifier, (2) cosine similarity between SBERT embeddings of original and generated text, (3) change in Flesch–Kincaid readability grade.

3 Results

3.1 Stylometric Differences

Stylometric analysis reveals clear differences between the two authors. Austen tends to produce longer, more structured sentences with higher readability stability, while Twain shows greater variability, more dialogue, and higher rates of expressive punctuation.

3.2 Embedding Space Visualisation

The UMAP projection of sentence embeddings shows two well-defined clusters corresponding to the two authors, with limited overlap. This suggests that transformer embeddings encode stylistic signals beyond semantics.

3.3 TF–IDF Lexical Markers

TF–IDF weights reveal stylistic tendencies at the surface lexical level. Table 1 summarises a subset of the strongest markers.

3.4 Style Transfer Evaluation

Style transfer results indicate that the style classifier assigns higher target-author probabilities to generated passages than to the original texts, but the absolute style scores remain relatively low (below 0.25 on average), suggesting only a partial shift in style.

Table 1 Top TF–IDF lexical markers for each author (logistic regression coefficients).

Author	Marker	Weight
Mark Twain	`tom`	4.01
Mark Twain	`king`	3.63
Mark Twain	`got`	3.21
Jane Austen	`mr`	-5.41
Jane Austen	`mrs`	-4.30
Jane Austen	`miss`	-3.89

Cosine similarity values around 0.55 (Austen→Twain) and 0.46 (Twain→Austen) indicate moderate preservation of the original semantic content. The negative Flesch–Kincaid deltas in both directions show that the generated passages tend to be stylistically simpler than the originals, often closer to a more neutral or modern register rather than a faithful reproduction of the original literary complexity.

Table 2 Style transfer evaluation summary (mean scores over 25 passages per direction).

Direction	Style score ↑	Cosine ↑	FK Δ
Austen → Twain	0.24	0.55	−2.22
Twain → Austen	0.09	0.46	−2.82

4 Discussion

The results support our first research question by showing that Austen and Twain form clearly distinguishable distributions in both stylometric feature space and transformer embedding space. This suggests that modern sentence embeddings are sensitive not only to semantic content but also to stable stylistic patterns present across an author's work.

Regarding our second research question, we observe that classical stylometry and TF–IDF markers provide complementary insights to embeddings. Stylometric features and lexical markers are easier to interpret and relate directly to traditional literary notions such as sentence rhythm, vocabulary richness, and dialogue frequency. Embeddings, in contrast, offer a compact representation that enables robust classification and smooth geometric visualisations of style.

For the third research question, the style transfer experiments demonstrate that prompt-based LLMs can reproduce several target-style characteristics while broadly preserving content. However, fine-grained stylistic control remains challenging: some outputs are overly generic, and content can be partially lost or rephrased in unexpected ways. The trade-off between stronger stylistic shift and higher semantic distortion is evident in the joint behaviour of style score and cosine similarity.

5 Conclusion

We presented a case study on literary style analysis and transfer between Jane Austen and Mark Twain, combining stylometric features, TF–IDF markers, sentence embeddings, and prompt-based generation with a large language model. Our findings show that stylistic signatures of the two authors are consistently captured by both classical and neural representations and that LLMs can perform meaningful style transfer in a zero-shot fashion.

Nevertheless, our work has several limitations: we focus on only two authors, we rely on automatic metrics rather than human judgments, and we do not attempt to disentangle style and content explicitly. Future work will extend the analysis to additional authors and genres, incorporate human evaluation of stylistic fidelity and content preservation, and explore more principled methods for controlling style within generative models.

Declarations

Data availability

The processed passages, stylometric features, and evaluation outputs used in this study are available in the accompanying project repository submitted with this report.

Code availability

All code used to preprocess the corpus, train the models, and generate the results is available in the accompanying project repository.

Author contribution

The project was conducted by a single author. All stages of the work (data collection, implementation, experiments, analysis, and writing) were carried out by the same author.

AI usage disclosure. The author used ChatGPT 5 as an assistive tool during the development of this project. ChatGPT was used to help write and refine parts of the code and to clarify technical concepts when needed. All code produced with AI assistance was reviewed, adapted, and tested by the author to ensure full understanding of the underlying methods and correctness of the implementation.

ChatGPT was also used to improve the clarity and structure of the written text, while all experimental design, data processing decisions, model training, evaluation procedures, and interpretation of the results were carried out by the author.

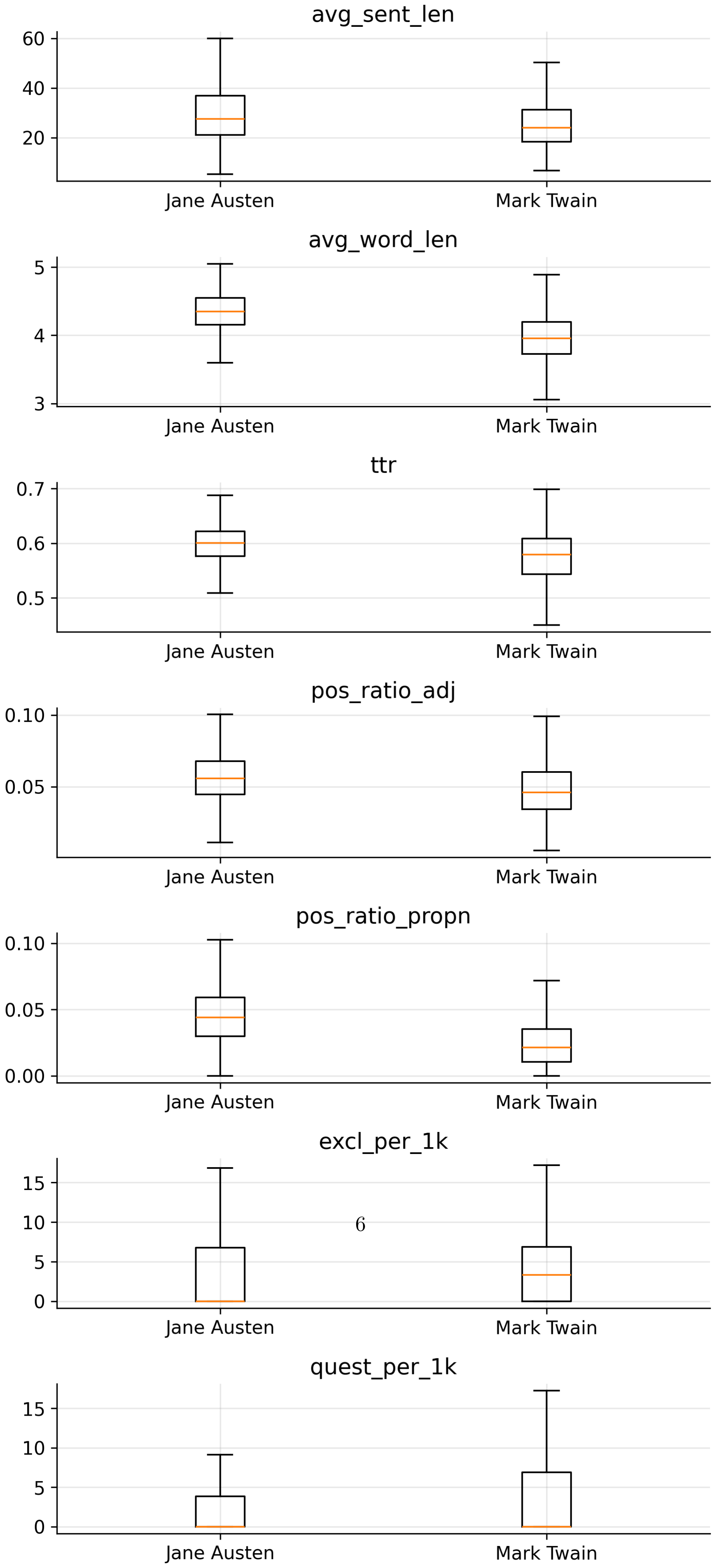

Fig. 1 Stylometric feature distributions for Austen and Twain passages.

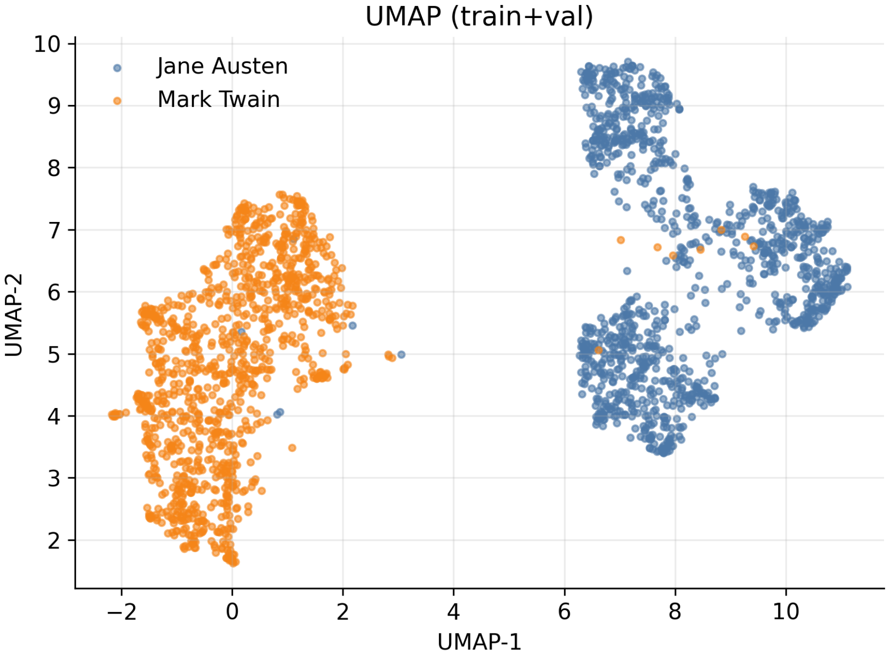

Fig. 2 UMAP projection of sentence embeddings coloured by author.