Project Report

Fine-Tuning a Language Model to Imitate Shakespeare's Writing Style

Angel Gantzia, Dido Stoikou, Vincent Tian

Abstract

Is it possible to generate structured data that quantitatively measures how well generative models can replicate an author's distinctive writing style? This inquiry extends beyond mere text generation to the nuanced task of producing text that captures the depth, style, and subtleties of Shakespearean English, providing a quantitative benchmark for the fidelity of these models to the original texts.

Introduction:

William Shakespeare is an emblematic author whose language has transcended centuries. This project dissects and replicates the linguistic essence of Shakespeare. Specifically, we will enlist various generative AI methods aiming to emulate William Shakespeare's iconic writing style. We utilize a dataset that has Shakespeare's quotes, and how they are translated in simple English. Central to our approach is the rigorous evaluation of how well generative, large language models capture the stylistic and thematic essence of Shakespearean English. We will transition from qualitative assessments that require human oversight to a structured automatic, quantifiable evaluation methodology.

By fine-tuning large language models using a dataset of Shakespearean texts and their modern translations, we aim to quantify stylistic fidelity and thematic coherence. This structured approach will allow us to objectively measure and enhance the capabilities of AI in reproducing complex literary styles. The results will not only advance our understanding of computational linguistics but also offer new tools for educational and creative endeavors.

We use two datasets in our study: the Shakespeare author imitation dataset (Xu et al., 2012) and the Shakescleare dataset. The former consists of 48,000 training sentences that mimic the style of William Shakespeare, contrasting with their modernized versions, while the latter contains 37,000 sentences directly from Shakespeare's plays and their modern English translations. These datasets highlight the language differences in vocabulary and syntax between Early Modern English and contemporary English, presenting the data in a dataframe format that neatly organizes Shakespearean English alongside its modern translations. This dual-layer provides a comparative framework essential for training our models to recognize and generate Shakespeare's distinctive linguistic features. By structuring this data, we ensure a comprehensive representation of stylistic nuances, thematic depth, and linguistic complexity inherent in Shakespearean English, laying a robust foundation for our project.

Background:

The burgeoning capabilities of large language models have encouraged the need to find robust evaluation frameworks that can quantify their performance and fidelity to specific linguistic styles or constraints. Specifically, the literature engages in stylistic dimensions such as sentiment, subjectivity bias, politeness, offensiveness, formality, genre based on audience (expert/layman). In the realm of evaluating the stylometric alignment of generated text, recent studies have developed innovative methodologies to tackle this complex challenge.

The study by Krishna et al. (2020) offers a nuanced approach to evaluating style transfer models, particularly focusing on multilingual formality transfer. It highlights the necessity of distinguishing between content preservation and stylistic transformation, setting a precedent for subsequent research in the field. Among other datasets, the authors also evaluate emulation of Shakespearean English in binary format.

Building upon the principles of style adherence and content fidelity, another study titled "How BERT Speaks Shakespearean English?" by Cuscito et al. (2024) provides a methodological leap by assessing historical bias in contextual language models. The authors introduce temporal valence scoring, assigning scores to words and phrases based on their alignment with a specific historical period of English. This method allows for a more granular evaluation of the models' ability to generate text stylistically and temporally aligns with a targeted linguistic epoch, such as Shakespearean English.

These aforementioned studies collectively inform the current discourse on LLM evaluation on emulating Shakespeare. These metrics are essential for pushing the boundaries of text generation to a level of sophistication that approaches human-like creativity and sensitivity to contextual nuance.

Methods:

I. Model Training and Imitation:

This project utilized two LLM architectures for training: GPT-2 and DaVinci. Training data for both models was drawn from the Shakescleare dataset, consisting of modern English passages alongside their Shakespearean translations. Due to resource constraints, a subset of 1,000 rows out of the total 48,000 rows was employed for training purposes. Moreover, in this project, we utilized the output produced by the Style Transformer model (Reformulating Unsupervised Style Transfer as Paraphrase Generation, by Kalpesh Krishna, John Wieting, Mohit Iyyer).

GPT-2, a transformer-based autoregressive model, operates by predicting the subsequent word in a sequence based on preceding words. For fine-tuning with GPT-2, a specialized formatting approach was adopted. A new column was introduced in the dataset, structured as follows: <s> (start token) + English translation + </s> (end token) + >>>> + (start token for Shakespeare translation) + Shakespearean translation + (end token for Shakespeare translation). This concatenated column served as input for GPT-2 fine-tuning. Subsequently, the values within this column were concatenated to create a .txt file for GPT-2 training. Standard GPT-2 tokenizer and pretrained model were utilized, with training parameters set to 3 epochs and a batch size of 8. During testing, input data was formatted as <s> + English input + </s>, which was then fed into the fine-tuned GPT-2 model to generate predictions. Extracting the model output involved retrieving the prediction between the and tags.

DaVinci, our second large language model (LLM), was also fine-tuned specifically for this project. Given its inherent prompt-completion functionality, we adopted a unique dataset formatting approach. The fine-tuning dataset consisted of prompts structured as follows: "prompt: System role: You are an expert author on Shakespeare. Write the following quote like how Shakespeare would say it:", followed by the English translation. The completion section was formatted as: "completion: This is how Shakespeare would say it:", followed by the Shakespearean translation. Employing the base DaVinci model, we generated model outputs using a process similar to that of GPT-2. Additionally, a temperature value of 0.2 was applied during inference to alleviate excessive variance in the output.

The Style Transformer model performs style transfer on text by maintaining the original meaning while changing the text's stylistic elements. This is achieved through a controlled paraphrase generation approach known as Style Transfer via Paraphrasing (STRAP), which creates pseudo-parallel data and uses inverse models to revert to the original style after transformations. The model is built on the transformers library (Wolf et al., 2019) using an encoder-free sequence-to-sequence framework (Wolf et al., 2018), incorporating segment embeddings and fine-tuning on GPT2-large. We retrieved the outputs of testing this model with the mentioned dataset, using different p-values for nucleus sampling (p=0.0 and p=0.6) to examine variations in style transfer performance. The output for the former is signified by the name “Style Transformer A”, while correspondingly the output for the latter is named “Style Transformer B”, as they will both be outlined in the results later.

II. Evaluation and Metrics:

To evaluate the efficacy of language models in reproducing Shakespearean English, we employ an evaluation framework that integrates four key metrics: the Binary Classifier, Content Preservation, Style Approximation, and Fluency. The Binary Classifier serves as the initial filter, determining if the model’s output possesses the distinctive stylistic features that characterize Shakespearean English. Following that, Content Preservation measures how faithfully the model retains the original message and essential details through the stylistic transformation. Next, Style Approximation assesses how closely the model’s output adheres to the unique stylistic elements that define Shakespeare’s writing, encompassing his metaphorical language, rhythm, and poetic devices. Lastly, Fluency examines the output's grammatical coherence and naturalness. Together, these metrics provide a comprehensive assessment of the model's ability to not only mimic Shakespeare’s style but to do so while preserving the integrity and fluidity of the original content.

- Binary Classifier: In our project, we use a Binary Classifier, specifically a fine-tuned DistilBERT-base-uncased model, to determine if the style of the generated text aligns with Shakespearean English and to evaluate the success of style transfers. This classifier is trained to detect whether a transformed sentence matches the targeted stylistic characteristics of Shakespearean English, with its effectiveness measured by its accuracy. DistilBERT, a streamlined version of BERT, is optimized for speed by being pretrained without human-labeled data, focusing on matching the BERT base model's output probabilities, predicting masked words, and minimizing the cosine distance between its hidden states and those of the BERT base model to retain language understanding. The classifier we use can be found here.
- Content Preservation: A major objective is to ascertain how well the essential content and meaning of the original Shakespearean work are maintained in the model-generated text.

- *BLEU Score* (Papineni et al.): A metric originally designed for machine translation, it measures precision of generated text by comparing it to reference Shakespeare texts. It counts matching sequences of words (1-gram up to 4-gram) between the generated text and reference. It penalizes shorter generated texts to avoid undersized outputs. Its scale is from 0 to 1, where higher values indicating better quality.

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}.$$

Then,

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right).$$

- *ROUGE-N Score* (Lin): Originally developed for summarization tasks, this metric measures recall, evaluating how much of the content in the reference texts appears in the generated text. It counts overlap of N-gram sequences between the generated and the original Shakespearean quotes. Its scale is from 0 to 1, where higher values indicating better quality.

$$\text{ROUGE-N} = \frac{\sum_{S \in \{ReferemceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count(gram_n)}$$

- Style Approximation: This set of metrics evaluates how closely the generated text imitates the linguistic style of Shakespearean English.
  - *Cosine Similarity* (Pennington et al.): With GloVe embeddings, using a 300-dimensional vector space built from a vocabulary of 42 billion tokens. By measuring the cosine of the angle between their GloVe vector representations of texts, this metric determines how closely the meaning of the generated text matches the original. Its scale is from 0 to 1, where higher values indicating better quality.

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

  - *Jaccard Similarity*: Measures the proportion of shared words to total unique words in both the original and the generated text, providing a straightforward content overlap metric to indicates the degree to which the distinctive lexical choices of Shakespeare are emulated. Its scale is from 0 to 1, where higher values indicating better quality.

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

- Fluency: The focus here is on the smoothness and naturalness of the text's flow, as well as its originality.
    - *PINC Score (Paraphrase In N-gram Changes)* (Chen and Dolan): Unlike traditional metrics, PINC assesses the novelty and structure of the generated text by examining the changes in N-gram patterns. It is a nuanced measure that penalizes over-repetition and rewards syntactic variety and complexity, thereby gauging the text's fluency and its resemblance to the creative and dynamic use of language found in Shakespeare's plays. It scales from 0 to 1, where lower values indicating better quality, with no novelty in the generated text; it's exactly the same as the candidate text with regard to the n-gram patterns being evaluated.

$$PINC(s,c) = \frac{1}{N}\sum_{n=1}^{N} \quad 1 - \frac{|\ \text{n-gram}_s \cap \text{n-gram}_c\ |}{|\ \text{n-gram}_c\ |}$$

Collectively, these metrics provide a comprehensive assessment framework, ensuring the generated text is not only thematically and stylistically akin to Shakespeare's works but also retains the poetic fluidity and inventiveness for which his writings are renowned.

Results:

Model Output	**Style Classifier**	**BLEU Score**	**Rouge-N Score***	**Cosine Similarity**	**Jaccard Similarity**	**PINC Score**
GPT-2	53.21%	6.74%	14.52%	83.29%	19.76%	85.61%
GPT-DaVinci	**79.69%**	5.59%	19.59%	87.74%	23.56%	80.88%
Style Transformer (A)	28.66%	7.21%	**22.91%**	**93.45%**	**36.38%**	**76.66%**
Style Transformer (B)	34.23%	**7.32%**	20.91%	93.30%	34.40%	78.20%

*Recall was used

The table above provides an overview of the average performance metrics derived from a test set consisting of 1461 sentences. In terms of the Style Classifier, GPT-DaVinci emerged as the top performer with an accuracy of 79.69%, closely trailed by GPT-2. Surprisingly, the two Style Transformers displayed the least proficiency in this aspect. However, a different trend unfolds when examining the other metrics. With the exception of the BLEU Score, where Style Transformer B exhibited the strongest performance, Style Transformer A consistently outshone the other models across metrics such as Rouge-N Score, Cosine Similarity, Jaccard Similarity, and PINC Score. Notably, Style Transformer B demonstrated slightly inferior performance across these metrics as well, indicating that both Style Transformers are notably adept at capturing Shakespeare's depth, style, and subtleties compared to the LLM models.

Example 1

English	My generosity to you is limitless as the sea, and my love is as deep
Shakespeare	My bounty is boundless as the sea, My love as deep
GPT-2	My goodness to thee is as infinite as the sea, And my love as deep
GPT-DaVinci	My love to you is sound, sans crack or flaw
Style Transformer A	I'm boundless, love is deep
Style Transformer B	I'm boundless, love is deep

Example 2

English	I'll frown at them as they pass by, and they can react however they want
Shakespeare	I will frown as I pass by, and let them take it as they list
GPT-2	I'll frown at them as they pass by, And they may be as they will
GPT-DaVinci	I'll set upon them as they pass along
Style Transformer A	I'll frown and let them take it as they list it
Style Transformer B	I'll frown and let them take it as a listing

Example 3

English	Please, sir, have patiencem
Shakespeare	I do beseech you, sir, have patience
GPT-2	I pray you, sir, patience, I pray you sir, patience, I pray you
GPT-DaVinci	Pray you, sir, have patience
Style Transformer A	I beg your, sir, be patient
Style Transformer B	I beg your, sir, be patient

Discussions:

The results of our project indicate a complex landscape regarding the ability of large language models (LLMs) to replicate the intricate style of Shakespearean English. GPT-DaVinci stands out in terms of style classification accuracy at 79.69%, suggesting its superior capacity for capturing the stylistic nuances of Shakespeare when compared to other models, including GPT-2, which scored significantly lower at 53.21%. However, the performance of the style transformers, particularly Style Transformer A and Style Transformer B, was less proficient in this regard, achieving only 28.66% and 34.23%

respectively. This lower performance in style classification indicates a potential gap in these models' ability to consistently identify or replicate the subtler stylistic elements that define Shakespearean English.

In terms of other metrics such as BLEU Score, Rouge-N Score, Cosine Similarity, Jaccard Similarity, and PINC Score, Style Transformer A consistently performed well, surpassing the other models in most metrics except for BLEU Score, where it was marginally outperformed by Style Transformer B. This suggests that while Style Transformer A might struggle with style classification accuracy, it excels in capturing the thematic and content-related aspects of the text, as evidenced by high scores in Rouge-N and Jaccard Similairty, which measure content overlap and shared lexical choices, respectively.

Interestingly, the Cosine Similarity and PINC scores highlight a nuanced understanding of Shakespearean language structure and innovative text generation by Style Transformer A, suggesting that the model can generate text that is both close to the original in meaning and stylistically varied. This could be indicative of the model's ability to balance content preservation with creative text generation, a crucial aspect in style transfer tasks.

The variances seen across different models and metrics underscore the challenge of optimizing a model that can excel at both style imitation and content preservation. While GPT-DaVinci and GPT-2 show robust style alignment, they may not fully capture the depth and thematic richness of Shakespeare's texts as effectively as the Style Transformer. This reflects the inherent trade-offs present in model design and training objectives.

The examples provided above illustrate the varying capabilities of different models in replicating Shakespearean English. For instance, in the first example, GPT-DaVinci and GPT-2 retain the poetic depth of the original text, whereas Style Transformers A and B simplify the phrase significantly, which likely contributes to their lower style classifier scores. In the second example, GPT-DaVinci adeptly captures the dynamic expression, outperforming the other models that offer plainer translations. The third example shows all models maintaining the core sentiment of patience, but with varying degrees of formal respectfulness, reflecting the nuanced challenge of preserving both the stylistic and thematic essence of the original Shakespearean text. These examples underscore the varied effectiveness of different models in handling the linguistic subtleties of Shakespearean English, reflecting the challenges inherent in automated style transfer where maintaining stylistic integrity along with the original content's essence is crucial.

In conclusion, these findings illustrate the inherent complexities in designing AI models capable of mimicking specific literary styles. While some models may excel in recognizing and replicating stylistic elements, others may better preserve the thematic and content integrity of the original texts. Future work will need to focus on enhancing model sensitivity to the subtleties of language style and content, potentially through more targeted training datasets, refined model architectures, or advanced training methodologies that can better encompass the multifaceted characteristics of Shakespearean English.

References:

Krishna, Wieting, and Iyyer. *Reformulating Unsupervised Style Transfer as Paraphrase Generation.* 2020. *https://arxiv.org/abs/2010.05700.*

Chen, David L., and William B. Dolan. *Collecting Highly Parallel Data for Paraphrase Evaluation*. 2016. *https://www.cs.utexas.edu/users/ml/papers/chen.acl11.pdf*, https://www.cs.utexas.edu/users/ml/papers/chen.acl11.pdf.

Cuscito, Miriam, et al. *How BERT Speaks Shakespearean English? Evaluating Historical Bias in Contextual Language Models*. 2024. *https://arxiv.org/abs/2402.05034*, https://arxiv.org/pdf/2402.05034.

Krishna, Kalpesh, et al. *Reformulating Unsupervised Style Transfer as Paraphrase Generation*. 2020. *https://arxiv.org/abs/2010.05700*, https://aclanthology.org/2020.emnlp-main.55.pdf.

Lin, Chin-Yew. *ROUGE: A Package for Automatic Evaluation of Summaries*. 2004. *https://aclanthology.org/W04-1013.pdf*, https://aclanthology.org/W04-1013.pdf.

Papineni, Kishore, et al. *BLEU: a Method for Automatic Evaluation of Machine Translation*. 2002. *https://aclanthology.org/P02-1040.pdf*, https://aclanthology.org/P02-1040.pdf.

Pennington, Jeffrey, et al. *GloVe: Global Vectors for Word Representation*. 2014. *https://nlp.stanford.edu/projects/glove/*, https://nlp.stanford.edu/pubs/glove.pdf.