Historical Text Generation using GPT2 Model

Wesnahika B. Charles
Tulane University, CMPS6730
Wcharles@tulane.edu

Abstract

This project presents a GPT-2 styled fine-tuned text generation model that is designed to output historical data and text based on metadata prompts such as year and region. Using carefully selected data, this model was trained in formatting entries which consisted of contextual metadata and actual content. The data was then preprocessed by Hugging Face's GPT-2 tokenizer and the training was done over multiple epochs with casual language modeling using both training and evaluation splits.

After training, the model was evaluated by multiple metrics. For example, BLEU and ROUGE for lexical overlaps, as well as BERTScore. When using these metrics, we note that the BLEU and ROUGE scores were quite low as there were low level surface level similarities between the texts, however, the BERTScore was quite moderate (at ~0.76) which would suggest that the model captured underlying historical themes and structure.

1 Introduction

In recent years, especially with the advent and popularity GPT-2 models, they have shown remarkable capabilities in generating fluent text and responses across a very vast range of topics and domains. However, for many of these language models, their effectiveness in very niche areas such as historical texts (where tone, structure, and factual content come into play) remain quite unexplored.

This project experimentally addresses this gap as the overall goal is to create a text generation model that, when inputted by the user, generates historical texts based on the year, and region/location. The output of this model is designed to showcase a coherent and fluent body of text that is relevant to the topic at hand and matches with the proposed input.

This will be done by collecting vast amounts of historical data (mainly primary sources, as well as literature) and using a pre-trained GPT-2 model to create a generative system capable of producing plausible, stylistically consistent historical narratives based on structured prompts which consists of metadata (Year and Region).

Beyond model training, the model incorporates evaluation through both traditional NLP metrics and semantic scoring.

2 Related Work

Natural language processing has been explored in the context of historical texts, particularly in tasks like modernizing archaic language or enhancing text accessibility for historians (Piotrowski, 2012). Although not focused on text generation, this work highlighted the challenges of handling historical language which is directly related to this project and some of its own limitations.

LSTM-based models have been applied to historical datasets to analyze language evolution, with their pattern recognition capabilities allowing them to track vocabulary changes across time periods (Hussein & Savas). This might be useful insight concerning working with data that spans centuries.

Recent advancements in data-to-text generation have been driven by transformers. One model improved generation accuracy by modifying the input encoding process (Gong et al), while another demonstrated that combining encoder models like BERT with GPT-style transformers enhances both coherence and contextual relevance in generated output (Chen et al, 2024).

Finally, the HuggingFace Transformers library has become a foundational tool in NLP research and applications, providing open-source access to state of the art models for text generation, classification and summarization (Wolf et al., 2020). This project directly builds on that ecosystem, leveraging it for fine tuning and deployment.

3 Methods

Regarding data collection, the original priority was to gather primary sources from different points of human history as well as make certain that they come from diverse references as well. One of the methods that was attempted was to scrape websites and databases that contained free and open-source literature and primary documents. This was done for Chronicling America (which consisted of American newspaper articles) as well as Project Gutenberg (contains primary sources and literature). The BeautifulSoup and Requests library was used, by skimming the titles, descriptions and (worst case scenario) the text, enough metadata was extracted for the Year, and with the use of a predefined python dictionary, the region was able to be found, categorized and placed in the structured dataset. Another method included using already existing python libraries such as Wikipedia or InternetArchive to collect articles and information based on category, key words and collections. That method was the most fruitful considering how I was able to collect straightforward, already available data without having to parse through HTML code and formatting.

To collect this data a “loader” script for each of my sources collected and formatted texts into data frames whose columns consisted of Title, Year, Region, and Source. These loaders were then all fed into a “Pipeline” script that ran each loader and combined all the data frames into one large dataset. The full dataset alone contained about 800 records.

When training the model, the dataset was split into training and evaluation sets, and a custom PyTorch class was used to tokenize the text using HuggingFace GPT-2’s tokenizer. The text was then segmented into fixed sized blocks of 1024 tokens and the maximum steps of 1000.

A DistilGPT-2 model was fine-tuned using the Hugging Face trainer with casual language modeling objectives. The training was then conducted across 8 epochs using a small batch size and learning parameters suitable for resource-constrained environments.

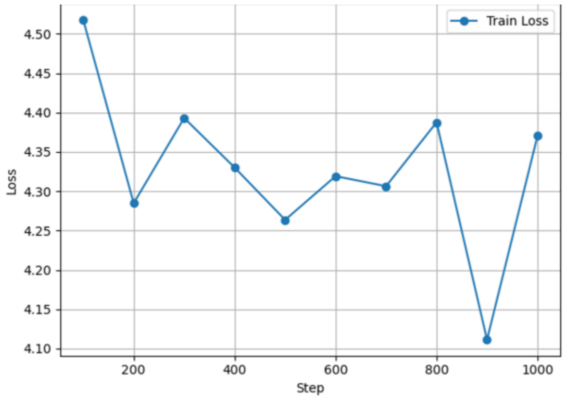

Figure 1: Training Loss vs Steps

The training process took a total of 3.5 hours to complete.

The training loss fluctuated between 4.1 and 4.5 which is somewhat expected since some of the data may be quite noisy due to how some of the data was parsed from scrapping.

4 Results

After fine-tuning the DistilGPT-2 model on the historical dataset, a function was created to allow for some showcasing of the model itself. A script located in the model's notebook was implemented to allow for model to generate a piece of text that resembles a historical document relevant to that time and place. Since the dataset's Year column is a string value (to accommodate ranges and BC dates) a function was created that takes the users input and turns it into a negative number whenever a BC date is given. Then it looks through the dataset and finds the closest matching entrie(s) based on input and returns a formatted prompt, to which it loads the fine tune model, tokenizes the prompt and generates new text that continues from the historical metadata prompt.

An example would be what is generated when a user inputs 1800 for year and France for Region

"[YEAR: 1823] [REGION: France] [SOURCE: Project Gutenberg] The Bible is one of the most well-known religious texts in Latin America, founded by a French missionary named T. Gage.[3]. By now it has been accepted as canonized for noncanonical purposes; but according to historian John O'Dale,[4], only about 578 copies exist on Wikipedia alone.[5][6](www). It was published after Spanish independence from Spain and subsequently under Pope Francis (1989). However many have since disappeared without further comment due some misconceptions regarding this text.[7]:13 So far, no other translations can be found online at http://bookshelves.org/en_wikipedia/."

As we can see, the output is quite interesting, there are some mentions of the French, and relevant places like Latin America, however the mention of Spain is not quite relevant to the prompt itself. Some strange wording and the fact that there wasn't a Pope Francis in 1989 (at least under that name) lets me know that this model is outputting some irrelevant text. However, it is easy to read, and the sentence structures are intelligible and mostly grammatically sound. Aside from human evaluation, quantitative metrics showed, on average low and underperforming scores for BLEU, ROUGE scores. This is expected given the creative and open-ended nature of generative historical writing where many valid inputs may differ lexically from the original source material.

Metric	F1
BLEU	0.0000
ROUGE-1	0.0170
ROUGE-2	0.0030
ROUGE-L F1	0.0106

Table 1: lexicon-based metrics

In contrast, the BERTScore averaged 0.7591, indicating a moderate degree of semantic similarity between outputs and reference texts. This suggests that while the model does not produce exact phrases, it captures some of the underlying themes and meanings present in historical records.

When evaluating Perplexity, overall, it received a score of 38.42, which is also not unexpected since my data is quite heterogeneous (varying in style, spelling and structure).

5 Discussion

The result of this project demonstrates the potential of fine-tuning transformer-based language models for generating historically accurate text based on structured metadata prompts. While the model achieved a relatively high perplexity of 38.42 and low BLEU and ROUGE scores, a BERTScore F1 of 0.76 indicates that the generated context captured semantic meaning even if it diverged from the reference text in wording. This suggests that the model learned to generate thematically appropriate narratives rather than replicating exact phrasing.

This project faced notable limitations and challenges, for example, training the model on just 1000 steps, took about 3.5 hours which highlights resource constraints despite using a lightweight model. Additionally, gathering high-quality, structured historical data proved difficult, requiring significant manual effort to clean and organize sources. The small dataset and inherent variability in historical language further limited the model's ability to generalize. Despite these challenges, the model still produced coherent, sometimes contextually relevant outputs showing promise for future applications with larger models, improved datasets, and integrated human evaluation.

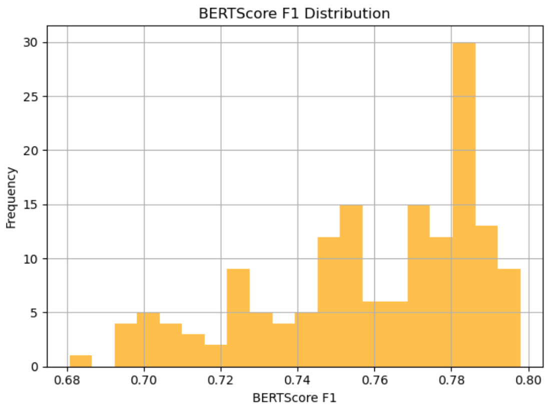

Figure 2: BERTScore F1 Distribution Histogram

Figure 1: A caption

6 Conclusion

This model explored the fine-tuning of a Distil-GPT-2 language model in order to generate historically themed text from structured prompts. Despite constraints related to dataset size, training time, and model capacity, the results demonstrate that even lightweight transformer models can produce semantically relevant and stylistically consistent historical narratives. While traditional evaluation metrics showed low surface-level similarity, BERTScore highlighted the model's ability to capture underlying meaning. Future work should be focused on scaling the dataset, leveraging larger models, and incorporating human feedback to further improve historical accuracy and fluency.

7 Division of Labor

This model and project were done in full by I, Wesnahika Charles. With me dividing my time to work on each aspect of the project as needed.

References

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

Piotrowski, Michael. "Natural Language Processing for Historical Texts." *Google Books*, Morgan & Claypool, 2012,

Hussein, Mustafa Abbas Hussein, and Serkan Savaş. "LSTM-Based Text Generation: A Study on Historical Datasets." *arXiv.Org*, 11 Mar. 2024, arxiv.org/abs/2403.07087.

Gong, Li, et al. "Enhanced Transformer Model for Data-to-Text Generation." *ACL Anthology*, aclanthology.org/D19-5615/. Accessed 15 Apr. 2025

Chen, Jiajing, et al. "A Combined Encoder and Transformer Approach for Coherent and HighQuality Text Generation." *arXiv.Org*, 19 Nov. 2024, arxiv.org/abs/2411.12157.

Wolf, Thomas, et al. "Huggingface's Transformers: State-of-the-Art Natural Language Processing." *arXiv.Org*, 14 July 2020, arxiv.org/abs/1910.03771.