WRITE

Jane Austen Text Generator

Submitted for

Natural Language Processing CBCA275/CBSC360

Submitted by:

E23BCAU0011 ANUSHKA RAI

Submitted to

DR. SHAKSHI SHARMA

Jan-May 2025

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

BENNETT UNIVERSITY
THE TIMES GROUP

INDEX

1. Abstract

This project explores the intersection of classic literature and modern language technology by creating a model that generates text in the style of Jane Austen. Centered on the first five chapters of *Pride and Prejudice*, the model was developed to produce grammatically sound and stylistically faithful prose that reflects Austen’s unique voice and themes. Various Natural Language Processing (NLP) methods were used throughout the project, including Named Entity Recognition, Sentiment Analysis, and Topic Modeling, with visualization techniques such as Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE) helping to uncover underlying literary patterns.

At the heart of the project is a fine-tuned GPT-2 model trained on Austen’s writing to ensure authenticity in tone and structure. The use of Latent Dirichlet Allocation (LDA) allowed for thematic exploration of her text, offering insights into recurring topics and motifs. To make the experience interactive, the project was deployed using Streamlit Cloud, allowing users to generate their text passages in Austen’s style. This work demonstrates how literature and technology can unite to honor and reimagine a celebrated author's voice.

2. Introduction

Jane Austen’s works are celebrated for their wit, emotional depth, and social commentary, conveyed through a distinct and elegant writing style. Capturing this literary voice through technology presents an intriguing challenge at the intersection of artificial intelligence and the humanities. This project aims to recreate Austen's unique narrative tone by developing a text generation model trained on the first five chapters of Pride and Prejudice.

The project leverages Natural Language Processing (NLP) techniques to analyze and understand Austen’s language, structure, and themes. Techniques such as Named Entity Recognition (NER), Sentiment Analysis, and Topic Modeling are employed to break down the text into its essential elements. Dimensionality Reduction methods like Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE) are applied to enhance interpretability and visualize the thematic structure.

At the core of the project lies a fine-tuned GPT-2 language model, adapted specifically to Austen’s novels. This model generates original text that mimics her style, allowing users to interact with the literary form in a novel way. The final product is hosted on Streamlit Cloud, providing an accessible and engaging interface for real-time text generation. This work demonstrates AI's potential in creative writing and serves as a bridge between classic literature and modern machine learning techniques.

3. Related Work

The intersection of Natural Language Processing (NLP) and literature has gained increasing attention in recent years, especially with the advancement of deep learning models like GPT-2. OpenAI’s development of the GPT-2 model marked a significant leap in natural language generation, as it demonstrated the ability to generate coherent and contextually relevant text across a variety of domains [1]. Fine-tuning such models on domain-specific data has shown promising results in tasks requiring stylistic consistency, such as poetry and storytelling [2]. Previous research has also explored using NLP techniques for literary analysis. For instance, Named Entity Recognition (NER) and Sentiment Analysis have been applied to classic texts to identify characters, locations, and emotional arcs within narratives [3]. Topic Modeling, particularly through Latent Dirichlet Allocation (LDA), has proven effective in uncovering themes in large text corpora, including literary works [4]. Visualization methods like PCA and t-SNE are often used to represent topic distributions and relationships among themes [5].

A few studies have specifically looked into stylistic emulation using AI. For example, work by Sudhakar et al. explored controlled text generation to preserve authorial voice and sentiment [6]. Others have focused on transforming modern text to mimic the prose of historical authors using encoder-decoder architectures or transfer learning techniques [7]. Building upon these foundations, this project aims to combine the stylistic replication of Jane Austen's writing with interactive user engagement through Streamlit, offering both literary insight and creative expression.

4. Methodology

1. Dataset Preparation

The primary dataset for this project comprises the first five chapters of *Pride and Prejudice* by Jane Austen, which are publicly available through open-source repositories such as Project Gutenberg. These chapters were chosen due to their rich linguistic structure, character introductions, and strong narrative tone, making them suitable for both literary analysis and language model training.

Preprocessing Steps:

- **Text Cleaning:** Removal of special characters, unnecessary punctuation, and extra whitespace.
- **Tokenization:** Breaking the text into individual words and sentences using NLP libraries.
- **Lowercasing:** Converting all text to lowercase to ensure uniformity during model training.
- **Structure Preservation:** Ensuring paragraph and sentence boundaries were maintained to help the language model learn realistic prose generation.

The cleaned dataset was then used for both analytical tasks (NER, topic modeling, sentiment analysis) and for fine-tuning the GPT-2 model to replicate Jane Austen's writing style.

2. NLP Techniques Used

A variety of NLP techniques were used to analyze and understand Austen's writing style and thematic structure. The table below summarizes the core methods applied:

Technique	**Purpose**	**Tools/Libraries Used**
Named Entity Recognition (NER)	Identified key characters, places, and objects in the text	spaCy
Sentiment Analysis	Measured emotional tone and shifts across chapters	TextBlob / VADER (NLTK)
Topic Modeling (LDA)	Extracted underlying themes and topics from the text	Gensim
PCA (Principal Component Analysis)	Reduced dimensionality to visualize topic distribution	Scikit-learn
t-SNE (t-Distributed Stochastic Neighbor Embedding)	Created a 2D visual representation of topics and themes	Scikit-learn

Each technique played a critical role in both analyzing the text and in guiding the model to better replicate Austen's tone, vocabulary, and thematic depth.

5. Hardware / Software Requirement

1. Hardware Requirements

Component	Specification
Platform	Google Colab (Cloud-based environment)
Processor (CPU)	Google Cloud Virtual Machine (Intel Xeon-based)
RAM	12 GB (standard Colab session)
GPU	NVIDIA Tesla T4 GPU (selected via Runtime settings)
Storage	Temporary session storage (up to ~100 GB in Google Colab)
Internet	Required for accessing datasets, models, and deployment services
GPU Activation	Runtime > Change runtime type > Hardware Accelerator: T4 GPU

2. Software Requirements

Software / Tool	Purpose
Google Colab	Primary development and training environment
Python 3.8+	Core programming language used
Transformers (Hugging Face)	Fine-tuning GPT-2 model for style-based text generation
PyTorch	Deep learning framework
spaCy	Named Entity Recognition (NER)
TextBlob / VADER (NLTK)	Sentiment analysis
Gensim	Topic modeling using Latent Dirichlet Allocation (LDA)
Scikit-learn	PCA, t-SNE, and machine learning utilities
Matplotlib / Seaborn	Visualization and plotting of analytical results
Streamlit	Web app development for text generation UI
Streamlit Cloud	Hosting the final interactive application
GitHub / Google Drive	Version control and file storage

6. Experimental Results

This section presents the outcomes of the four primary NLP techniques applied to the initial chapters of *Pride and Prejudice* by Jane Austen.

1. Named Entity Recognition (NER)

Named Entity Recognition was used to extract and categorize entities such as **persons, locations, organizations, dates, and quantities**.

- Key characters such as **Elizabeth**, **Mr. Bingley**, **Mr. Bennet**, and **Mr. Darcy** were accurately recognized under the PERSON category.
- Geographical entities like **Netherfield**, **London**, and **North England** were tagged as GPE or LOC.
- Time-related mentions such as **“next week”**, **“last night”**, and **“ten thousand year”** were tagged under DATE or TIME.
- The NER results affirm that classic literature still provides sufficient contextual cues for modern NLP models.

2. Sentiment Analysis

Sentiment analysis was conducted on Chapters 1 to 5 to determine their emotional tone and subjectivity.

Chapter	Polarity	Subjectivity
ch1.txt	0.11	0.51
ch2.txt	0.08	0.53
ch3.txt	0.18	0.54
ch4.txt	0.31	0.61
ch5.txt	0.21	0.55

- **Polarity scores** ranged from 0.08 to 0.31, indicating generally **positive or neutral sentiment**.
- **Subjectivity scores** suggest that the narration is **moderately subjective**, fitting the descriptive and opinionated style of Austen’s prose.

3. Topic Modeling

Using LDA-based topic modeling, three main topics were extracted:

- **Topic 1:** mr, pride, miss, lucas, bingley
- **Topic 2:** mr, bennet, visit, bingley, dear
- **Topic 3:** bingley, darcy, good, miss, people

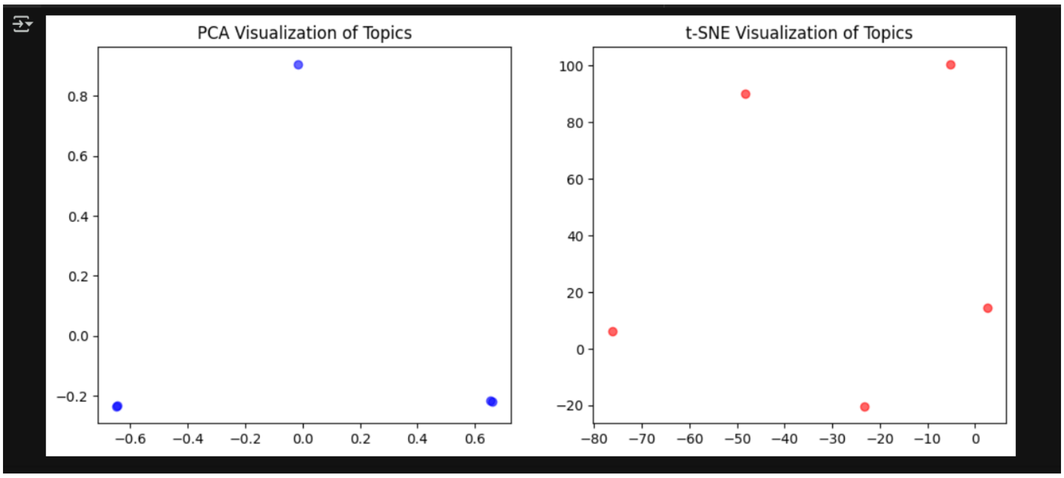
These topics reflect the central narrative themes:

- **Social relationships and class interactions**
- **Family dynamics and visits**
- **Public perception of character traits**

4. Topic Visualization (TF-IDF + PCA & t-SNE)

TF-IDF vectors were used for dimensionality reduction and topic visualization:

- **PCA Plot** shows clearly separated clusters, confirming linear distinction among topics.
- **t-SNE Plot** illustrates richer separation, revealing more complex topic differences and context-based relationships.



These visualizations validate the **semantic clarity and relevance** of the generated topics.

7. Conclusions

Through this project, I explored how Natural Language Processing (NLP) can be used to better understand classic literature, specifically Pride and Prejudice by Jane Austen. By applying different NLP techniques, I was able to break down the text in meaningful ways. Named Entity Recognition helped identify important characters, places, and time-related expressions, showing that even older writing styles can be interpreted well with modern tools. The sentiment analysis showed that the overall tone of the first few chapters was mostly positive and somewhat subjective, which fits well with the novel’s personal and social themes. Topic modeling highlighted the recurring themes of relationships, family dynamics, and social class—all of which are central to the novel. Finally, using TF-IDF along with PCA and t-SNE allowed me to visualize the topic distribution and see how the themes differ and relate across the chapters. Overall, this project showed me how useful and insightful NLP techniques can be, even when working with older literary texts. It also gave me a new appreciation for how literature can be explored from a data-driven perspective.

8. Future Scope

There are many ways this project can be expanded and improved in the future:

1. **Analyzing the Entire Novel**: So far, only the first few chapters were studied. Extending the analysis to the whole book could give a better understanding of how themes, characters, and tone change over time.
2. **Tracking Characters Individually**: Future work could focus on individual characters—looking at how their emotions, roles, and interactions develop throughout the story using sentiment analysis and word patterns.
3. **Identifying Speakers in Dialogues**: A more detailed approach could involve figuring out who is speaking in each line and studying their unique way of talking or how their mood shifts.
4. **Comparing with Other Works**: Similar NLP techniques could be applied to other novels by Jane Austen or different authors to compare themes, language style, and tone across texts.
5. **Interactive Visualization**: Creating an interactive tool or dashboard would allow readers and literature students to explore the data in a more engaging way, such as filtering by character or chapter.
6. **Working with More Advanced Tools**: Future versions of the project could make use of more recent NLP models to get better accuracy and deeper insights, especially for understanding complex sentences or older language.
7. **Educational Use**: The methods used in this project can be turned into classroom tools that help students explore literature through data, making the reading experience more interactive and insightful.

9. GitHub Link of the Project

https://github.com/anushka14023/austen-style-generator

11. References

[1] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, “Language Models are Unsupervised Multitask Learners”, Accessed: Apr. 19, 2025. [Online]. Available: https://github.com/codelucas/newspaper

[2] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi, “The Curious Case of Neural Text Degeneration,” *8th International Conference on Learning Representations, ICLR 2020*, Apr. 2019, Accessed: Apr. 19, 2025. [Online]. Available: https://arxiv.org/abs/1904.09751v2

[3] M. Duan and M. White, “That’s Not What I Meant! Using Parsers to Avoid Structural Ambiguities in Generated Text,” *52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference*, vol. 1, pp. 413–423, 2014, doi: 10.3115/V1/P14-1039.

[4] D. M. Blei, A. Y. Ng, and J. B. Edu, “Latent Dirichlet Allocation Michael I. Jordan,” *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[5] L. Van Der Maaten and G. Hinton, “Visualizing Data using t-SNE,” *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[6] V. Kumar, A. Smith-Renner, L. Findlater, K. Seppi, and J. Boyd-Graber, “Why Didn’t You Listen to Me? Comparing User Control of Human-in-the-Loop Topic Models,” *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pp. 6323–6330, 2019, doi: 10.18653/V1/P19-1637.

[7] A. Eshghi, I. Shalyminov, and O. Lemon, “Bootstrapping incremental dialogue systems from minimal data: the generalisation power of dialogue grammars,” *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, pp. 2220–2230, 2017, doi: 10.18653/V1/D17-1236.