Executive Summary

Content Creation Machine Learning Analysis Report

## Presented By:

Becca Roeth

Ethan Houser

Kyle Bell

Kyle Parks

Makayla Whipple

Chantal Borchardt

# I. Introduction Summary

We were tasked with building text generation models with the ability to replicate the writing style of Jane Austen, as well as other authors such as the King James version of the Bible, the Narnia series, and HG Wells. To achieve this, we employed a Recurrent Neural Network (RNN), which excels in recognizing intricate patterns and relationships within texts. Using this model, we are able to mimic the unique literary style that sets apart authors like Jane Austen.

We built four RNN models to capture the styles of each author, each one trained on a separate author. This way we can load each model and gather predictions for a given prompt. In the end, due to copyright concerns, we did not end up using the RNN model trained on the Narnia books.

# II. Addressing Questions

When it came to deciding whether to use LSTM layers or GRU layers in the network, we chose to focus on LSTM layers. When using a single GRU layer, the model mimicked full words and sentences, but it was unintelligible, and the words were misspelled to the point of being unrecognizable. We then switched to using a single LSTM layer, dramatically reducing the amount of incoherent words produced by the model.

## Question: Often when we are generating text we see something like this: *"we counter. He stutn co des. His stanted out one ofler that concossions and was to gearang reay Jotrets and with fre colt otf paitt thin wall. Which das stimn".* What would you recommend to improve our results? Based on your initial analysis of the data, your team feels:

### ***You should train for more iterations.***

Training for more iterations allows the model to go from producing character strings that look like words to producing real english words. However, training for too many iterations will cause the model to “overfit” and begin reproducing the training text verbatim.

### ***You should try a different kind of cell (for example, consider switching to/from an LSTM, GRU, etc.).***

Switching to an LSTM cell allowed our model to begin producing words in fewer training epochs (less time) than when using a GRU cell.

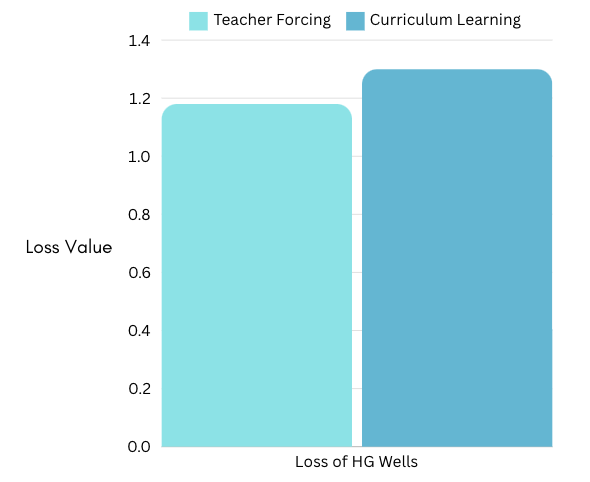
Something that would give us the most promise to a quality model would be Including more training data which makes it more coherent and better able to write a wider variety of texts, but putting in too many sources may limit how well it sounds like Jane Austen’s voice. Switching from character embedding to word embedding will also allow the model to learn word relationships, giving it a better chance to produce full, coherent sentences. If we combine these things, we would have a robust model that can generate coherent sentences in the style of people’s favorite authors.

## Question: I'm wondering what your views are on using a teacher forcing strategy compared to a curriculum learning strategy?

When it comes to a teacher forcing strategy it is the process of feeding back the model with observed sequence values after each step. This helps the model to not be influenced by a chain of incorrect predictions. Curriculum Learning is a training strategy inspired by how humans and animals learn incrementally from easier to more difficult tasks or examples. Rather than exposing the model to all training examples uniformly, curriculum learning starts with easier examples or simpler sequences and gradually increases the complexity or difficulty of examples as training progresses.

We tried a basic curriculum learning strategy, which feeds all previous epochs back into the model so it can learn from previous mistakes. The curriculum strategy got a loss score of 1.3 while the teacher strategy got a loss score of 1.18 as well as better quality text (*Fig 1*). The issue with our test is that curriculum learning requires more resources as it is less efficient. We would need to train the model for longer before we should see any real improvement.

*Fig 1*



## Question: Our previous team used logits in the output layer and then used Sparse Categorical Cross Entropy as the loss function. Are you planning to use that approach as well?

To help improve the models performance and enhance its understanding of basic language constructs the following recommendation can be made:

### ***We should not use any additional text, because it will change the style of the generated text.***

Adding additional text will improve the model's ability to generate words and sentences, but the more text that it is trained on, the less it will sound like Jane Austen. On the other hand, without a wide range of sources, the model may struggle to generate sensical text based on prompts that are not in the training works.

### ***We should only use works that are out of copyright and now in the public domain, such as Jane Austen, or other older works.***

This is the most ethical course of action, as it minimizes harm to authors who are being read for their unique ideas and style.

### ***We could use all of Wikipedia, or other creative commons works.***

Moral, but the plethora of content on Wikipedia will muddle the target style that we want the model to produce. Also, making content from Wikipedia usable would take a lot of effort and time. If the breadth of content from Wikipedia is ever deemed useful, the time it would take to clean the data might hold back future teams from using it. Wikipedia pages hold a lot of metadata and citations that one might not want to influence their language model. That is the case for Recurrent Style Neutral Networks for language. Learning Language Models are more flexible. They require little to no text reformatting.

### ***We could use all of the Internet, since we will not be copying any of the text verbatim.***

This is the approach taken by models such as ChatGPT, which was trained on datasets such as Common Crawl. However, this may not be moral, as it does not ask permission from the authors of the text of the internet. For the major part of the text on the internet, the author did not give permission for AI to use their text when that text was written. [As an example, Reddit may have sold its posts to Google’s “AI Overviews”, but the posters of Reddit did not make their posts with the intent that AI would use it.]  
Any text used to train an AI can eventually be repeated verbatim, or near enough to verbatim that it would qualify as plagiarism. [See, for example, [Scalable Extraction of Training Data from (Production) Language Models](https://arxiv.org/pdf/2311.17035)]

# 

# III. Data Exploration

Each dataset used are samplings of each Author’s bibliography with the exception of the King James Bible, which has a collection of many authors. They are in plain text and edited to only contain the content of the book. Various formatting quirks of the books have been edited out so as to not throw off the end results of the models. The training datasets used are all in the public domain: Jane Austen, HG Wells, and the King James Bible.

When it comes to embedding within machine learning models, embeddings are representations of values or objects like text, images, etc. that are designed for the finding relationships within data.

For our model we used the approach called character embedding which takes character-level relationships and encodes them into id’s. The model can then vectorize these, making them suitable for a neural network. This allows the model to recognize patterns and relationships between characters, then take in selected text and generate meaningful outputs.

# IV. Machine Learning Model

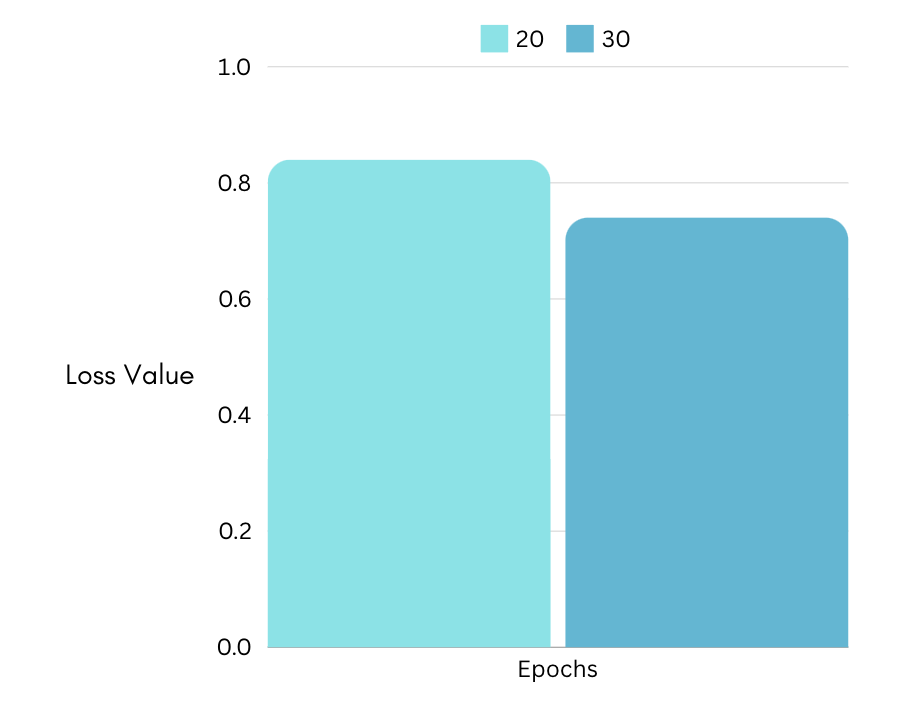
## How the model is built

Three models were trained to each dataset with the goal of getting something that generates original text in the style of said dataset. Training weights of the best version of each model were then saved for later use and analysis.

We took the basic model structure, which consisted of an embedding layer, GRU layer, and a dense (or output) layer, and replaced the GRU layer with a LSTM layer. We set the recurrent initializer to ‘glorot\_uniform’ which helps in setting the initial recurrent weights in a way that promotes stable and efficient training. We also set the forget gate bias to 1. There are multiple sources explaining its benefits, such as helping to mitigate the problem of a vanishing gradient, but we found that it helped our model to produce fewer typos in its word generation.

After training our model for 20 epochs, the loss of the model [sparse categorical cross entropy] was 0.84. The loss in this case measures how closely the model is “copying” the original text. We decided to train the model for another 10 epochs for a total of 30 epochs, bringing the loss down to 0.74 (*Fig 2*). The increase in epochs greatly helped to improve the model’s accuracy in generating words.

*Fig 2*



We also tested using a LSTM and GRU together. The loss of the LSTM/ GRU combo was also 0.84 after 20 epochs, but the generated content was gibberish. For this reason, we decided to use only the LSTM layer.

# V. Results, Action Items, and Limitations

With the starting input: “The world seemed like such a peaceful place until the magic tree was discovered in London.”

## Results

#### Jane Austen trained model:

* *“But Jane may there be a presumption for yourself, and I am not sorry to know that she would ever have been sometimes very much dislike; she would make you together, it ought to be impossible; for we must \_not\_ all Turnoun’s marriage, gives to the fortnight, it gives her guess?”*

#### King James Version of the Bible trained model:

* *“And Jesus pleased them unto upon the ground, and he believed the Chereshites were gathered James visels. Then Jethemiah called together all the heads which beginging him with him to prophesy in his secret. And when every one in Jerusalem saw in the priops could consume the power of the Lord, and to Jeremiah four God.”*

#### HG Wells trained model:

* *“All this time he felt to be instructed, and he still repeated to us and drink and forgotten the dis marchlopile of advanced by the talk hard and hirs, but long with themselves, way, Greham's moments, but these does not.”*

## Action Items & Conclusions

* The model that HackPressIO provided used:

from tensorflow.keras.layer.experimental import preprocessing

preprocessing.StringLookup(),

but in newer versions of TensorFlow, StringLookup is no longer experimental and should be imported using:

from tensorflow.keras.layers import StringLookup

* Switch to a word embedding model:
  + Switching to word embedding will enable the model to learn the relationships between words rather than characters, greatly increasing the model’s ability to generate coherent and meaningful sentences. Due to time constraints and difficulty understanding/ implementing the word embedding, we did not use it for our model here.

## Limitations

* We were able to train our model using the Narnia books by C.S. Lewis. However, due to copyright issues we can’t use this model. We initially believed that since Narnia is set in London and is well known for its magical elements that it would be perfectly suited for interpreting the phrase “The world seemed like such a peaceful place until the magic tree was discovered in London.”
* We do not currently have enough GPU space for multiple epochs which would reduce the need for large amounts of text and keeping the voice of the original author. Things like Curriculum Learning would be better for more rounds of training and we’ve already proven that more epochs would help.
* We do not currently have access to large amounts of text to train our models on. The more text that we can provide to our models, the more accurate they will be. The model will also lose the unique writing style of each author if it is trained on more text sources, so training on too much information would also cause adverse side effects.
* As mentioned in Action Items & Conclusions, word embedding would greatly increase our models accuracy. Due to time constraints, we chose to use character embedding for simplicity and efficiency.

# VI. Python Notebooks

Master Models:

* [Jane Austen](https://colab.research.google.com/drive/1CeV8gEHx1NwmZUCmViKj2J8M5NjWInhY?authuser=3)
* [KJV Bible](https://colab.research.google.com/drive/1hv1-n4jNZl68kBZY9f9toAsIcz6q_ShA?authuser=3)
* [HG Wells](https://colab.research.google.com/drive/1asTUhJnZ1ea71AfgVeRpI6tFNNgFOKKm?authuser=3)

Text files:

* [austen.txt](https://raw.githubusercontent.com/byui-cse/cse450-course/master/data/austen/austen.txt)
* [hgwells.txt](https://drive.google.com/file/d/1m0xj_drntNxursEwHxDEwVndGJxnzVhb/view?usp=drive_link)
* [KJV.txt](https://docs.google.com/document/u/3/d/1PoP5uvydsieeoGmNRhiKHKOWykO33CoLCEvp_KzEYF4/edit)