Module 6: Executive Summary

Presented by:

Joaquin Solis

Scott Townsend

Zach Holcomb

Isaac Weyland

I. Introduction, Data, and Insights

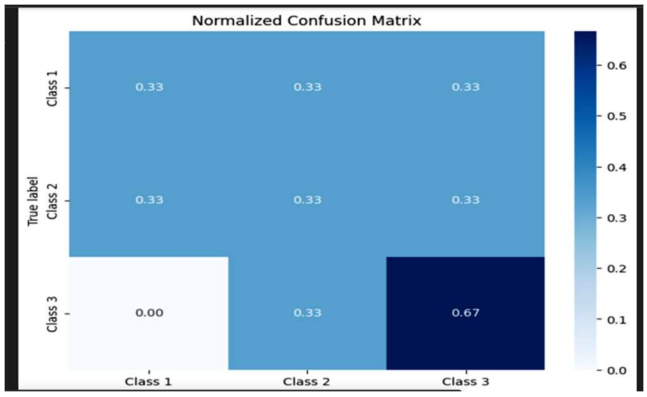
This project looks at how we can prepare and analyze written text from well-known public domain books to explore natural language processing (NLP). We used works by authors like Mark Twain, Edgar Allan Poe, and H.P. Lovecraft, which allowed us to examine a range of writing styles and word choices. These texts came from Project Gutenberg, a free online library of public domain books.

To get the data ready for analysis, we cleaned and organized the text by removing unnecessary characters, making everything lowercase, and separating punctuation from words. This helped ensure consistency and made the data easier for the model to understand. By looking at the cleaned-up data, we noticed patterns in how different authors write, like Austen's focus on conversations or Poe's detailed descriptions.

Key insights from the project include discovering Project Gutenberg as a valuable source for historic books, making it an excellent resource for training data. We also learned the importance of balancing input size for the model—too long, like the collected works of Poe, results in slow processing, while too short, like the Lovecraft book, lacks sufficient data for training. Additionally, since each writer has a unique style and vocabulary, we needed to tailor the data preparation process, adjusting text cleaning, punctuation handling, and text segmentation to fit these differences.

Confusion Matrix

In this confusion matrix, we can see detailed insights into the types of errors the model is making by showing how often predicted labels match actual labels by measuring predicted and true labels. The purpose of this confusion matrix is to identify strengths and weaknesses in the classifier's performance across individual classes and highlights which classes are frequently confused, allowing us to focus on model improvements for those areas. Below is a snipped of our confusion matrix:



The darker and lighter colors represent intensity or frequency of predictions for the classes. The darker colors indicate higher values (more predictions) while the lighter colors show the opposite. The shading is essentially showing us how well our model is creating our desired written texts. As we can see, the model is doing relatively well. There is one noticeable light spot

indicating low performance, but the other spots are much darker displaying how well the model is doing.

II. ML Model

Model Approach:
We built a simple machine learning model to turn the text into numbers that a computer can understand and analyze. Using tools like TensorFlow and Keras, we created a step-by-step process for breaking the text into smaller pieces, turning those pieces into numbers (a process called embedding), and then summarizing the information. This approach helped capture the important features of the text, like the meanings of words and their relationships.

Performance:
The model we created achieved significant performance, with the model creating relatively coherent sentences that flowed naturally and did as we were tasked. Despite this success, challenges such as long training times (some epochs taking 3-4 minutes) and overfitting persisted. Addressing these issues required careful tuning and applying data augmentation techniques.

Challenges:
While training the model, we faced some challenges, like how to handle sentences of different lengths or the unique ways different authors write. To solve this, we adjusted how we processed and grouped the data. Even with these challenges, the model worked well. It was able to pick up on patterns in the text and could be useful for tasks like figuring out which author wrote a piece of text or understanding the mood of a passage.

Best Use Case:
This model works best for jobs that need a basic understanding of language, such as summarizing text, finding the author of a book, or analyzing themes. In the future, we could improve the model by adding more data from Project Gutenberg, looking at other features like grammar, or trying out newer, more advanced models.

III. Python Notebooks

Notebook

IV. Discussion Responses

1. Do you think we should be using LSTM layers or GRU layers in this network?
LSTM layers are ideal for maintaining coherence over long sequences like a 1000-character text. However, if training time becomes a concern, experimenting with GRU layers is worth considering, as they are faster and often perform similarly.

2. 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?
B. Adding more layers could help the model capture richer patterns and improve text quality. You might also refine the training dataset and optimize hyperparameters for better results.

3. We're really trying to impress our investors with your work here. What would give us the most promise for both a quality model, but also something that could get people excited?
To balance quality and excitement, focus on a robust model architecture (like LSTM/GRU with attention mechanisms) and demonstrate clear, impactful outputs that highlight its capabilities, such as generating cohesive and creative text.

4. I'm wondering what your views are on using a teacher forcing strategy compared to a curriculum learning strategy.
Curriculum learning offers a stable approach by starting with simpler inputs and progressively increasing complexity, helping the model generalize better during inference. It's an excellent choice for incremental learning.

5. 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?
Yes, based on the previous team's approach, I would also recommend using logits in the output layer and Sparse Categorical Cross-Entropy as the loss function for your model. If it does not end up working, we could go with a different approach.

6. We've been making good progress with our target author, but we feel we could improve our model's performance by supplementing with additional text, to help it learn basic language constructs better.
Which of the following would you recommend?
B. We should only use works that are out of copyright and now in the public domain, such as Jane Austen, or other older works.

V. Summary

This case study explored the creation of a natural language processing (NLP) model using TensorFlow to process and analyze text from various authors. By downloading and pre-processing these texts, the model was trained to understand patterns, vocabulary, and language structures. The case study emphasizes the importance of cleaning data, customizing preprocessing steps for different authors, and efficiently managing computational resources during training. This work demonstrates how machine learning can provide valuable insights into literature and language. Our model was able to generate a coherent thought extracted from other texts.