2025

Author Identification with Long Short-Term Memory Neural Networks

PDAN8412 PART 1

MAXIMILIAN WALSH – ST10203070

30 SEPTEMBER 2025

Table of Contents

Introduction

This analysis aims to build a text classification model capable of identifying the author of short literary excerpts. This is a multi-class classification problem where the model must distinguish between three known gothic authors: Edgar Allan Poe (EAP), H.P. Lovecraft (HPL), and Mary Shelley (MWS). This can be seen as valuable to both literary scholarship, where stylistic analysis is essential, and broader applications of natural language processing (NLP) like plagiarism detection and forensic linguistics (OpenAI, 2025; Muller & Guido, 2016).

The dataset used, Spooky Author Identification, was sourced from Kaggle and contains nearly 20,000 text fragments labelled by author (Kaggle, 2017). After the cleaning and deduplication process, approximately 19,500 unique records were kept. The dataset is balanced across the three authors and captures diverse writing patterns within the Gothic genre, and is thus suited to this task (OpenAI, 2025; Kaggle, 2017).

The modelling process implements a full NLP pipeline from exploratory data analysis, to understand balance and text length distributions, to feature engineering for text preprocessing, tokenisation, padding, and integer sequence transformation, to finally the training of the Long Short-Term Memory (LSTM) neural network (Muller & Guido, 2016). LSTMs are well-suited for this task as they are effective at capturing sequential and contextual patterns in text (Muller & Guido, 2016).

The evaluation of the model was done through numerous metrics like accuracy, precision, recall, F1 score, ROC-AUC, confusion matrix, and learning curves. Providing a comprehensive and holistic overview of the modelling capabilities. After which, a retraining phase was conducted to combat overfitting and test model robustness through stronger regularisation and adaptive learning rate scheduling (OpenAI, 2025).

It can be said that the project showed feasibility and effectiveness in using deep learning methods for authorship attribution. Both challenges in overlapping gothic vocabularies and strengths of LSTM models in capturing literacy style were highlighted in the findings (OpenAI, 2025).

Dataset Justification

Source: Kaggle – “Spooky Author Identification” by Kaggle (Kaggle, 2017).

Size: 19,579 text fragments, 2 columns (text, author)

Target Variable: author (multi-class categorical: EAP, HPL, MWS)

The selected dataset is the Spooky Author Identification dataset from Kaggle, which contains nearly 20,000 text excerpts written by three famous authors: Edgar Allan Poe (EAP), Mary Shelley (MWS), and H.P. Lovecraft (HPL). The dataset chosen is well-suited for building a Long Short-Term Memory (LSTM) recurrent neural network for multi-class text classification for the following reasons:

- Relevance to Task: The dataset aligns directly with the project objective to build a model that can guess the author of a passage based on writing style. Each row is labelled with one of three authors, making this a supervised multi-class classification problem, a suitable task for RNN/LSTM models to capture sequential and stylistic patterns in language (OpenAI, 2025).
- High-Quality Labels: Explicit author labels (EAP, MWS, and HPL) are offered in the dataset, meaning there is no need to engineer or infer classes (Kaggle, 2017). Offering clarity and consistency in the training process and enabling the model to focus on stylistic signals like vocabulary, phrasing, and sentence structure (OpenAI, 2025).
- Sufficient Volume of Data: After the cleaning and duplicate removal process, the dataset offers 18,047 unique text excerpts. This amount comfortably meets the project subminimum of 10,000 entries. This large sample size supports the training of deep learning models and evaluating their ability to generalise well (Muller & Guido, 2016).
- Natural Language Sequences: Given that each entry is a raw text excerpt that is often a sentence or short paragraph, the data is inherently sequential, as words depend on previous words (Kaggle, 2017). This is a core strength of LSTM networks, which retain long-term dependencies and stylistic cues across sentences (Muller & Guido, 2016; OpenAI, 2025).
- Class Diversity: This dataset is relatively balanced between the three authors, with EAP at 7,044 entries, MWS at 5,552 entries, and HPL at 5,451 entries. Meaning the model learning is fair without excessive bias towards one class. Although some noise was present (invalid author labels), these entries were removed in pre-processing.

- Contextual Richness: The dataset contains distinct writing styles; EAP's gothic and dramatic tone, MWS's descriptive and romantic prose, and HPL's cosmic horror and archaic vocabulary (Kaggle, 2017). Such stylistic differences are suited for the RNN to learn author styles (OpenAI, 2025).

To conclude, the Spooky Author Identification dataset is considered an ideal choice because it's labelled, balanced, large enough, and text-sequential, making it appropriate for this LSTM-based recurrent neural network task.

Exploratory Data Analysis (EDA)

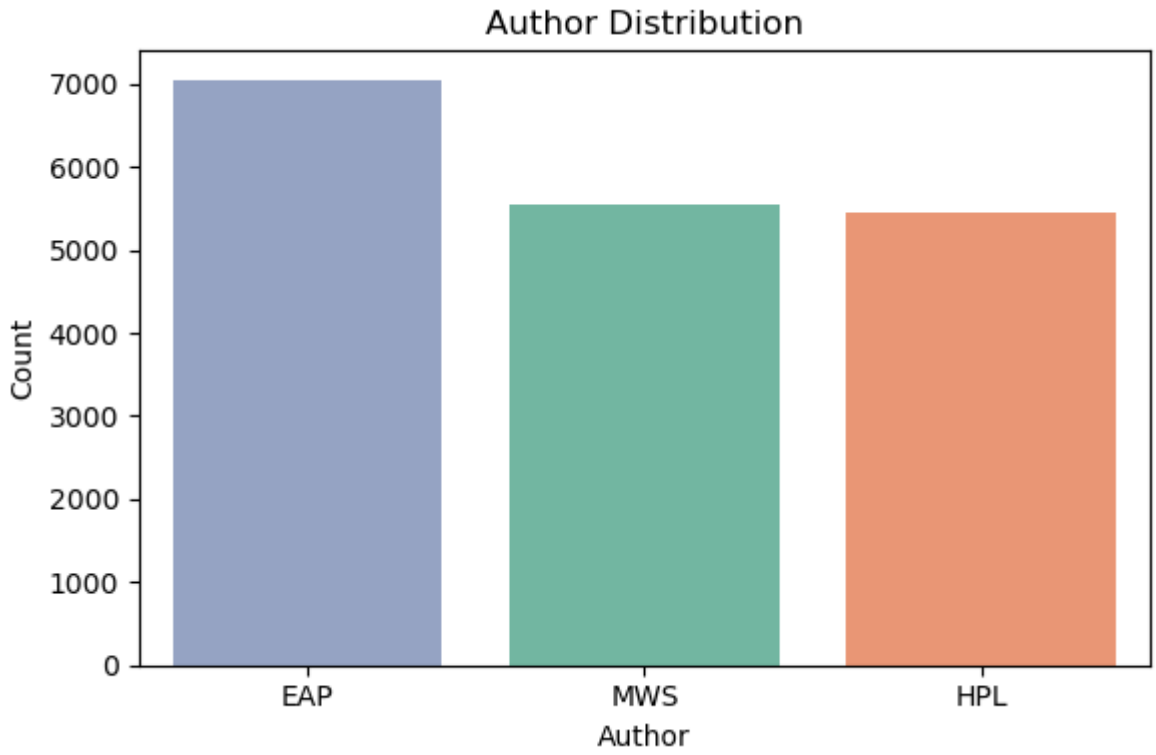

Figure 1: Author Distribution Bar Chart

The bar chart illustrated in Figure 1 shows a balanced distribution across the three different authors: EAP, MWS, and HPL. These results reduce the need for heavy resampling and justify this dataset for the task, as this balance is important for recurrent neural networks, which are sensitive to class imbalances (Muller & Guido, 2016).

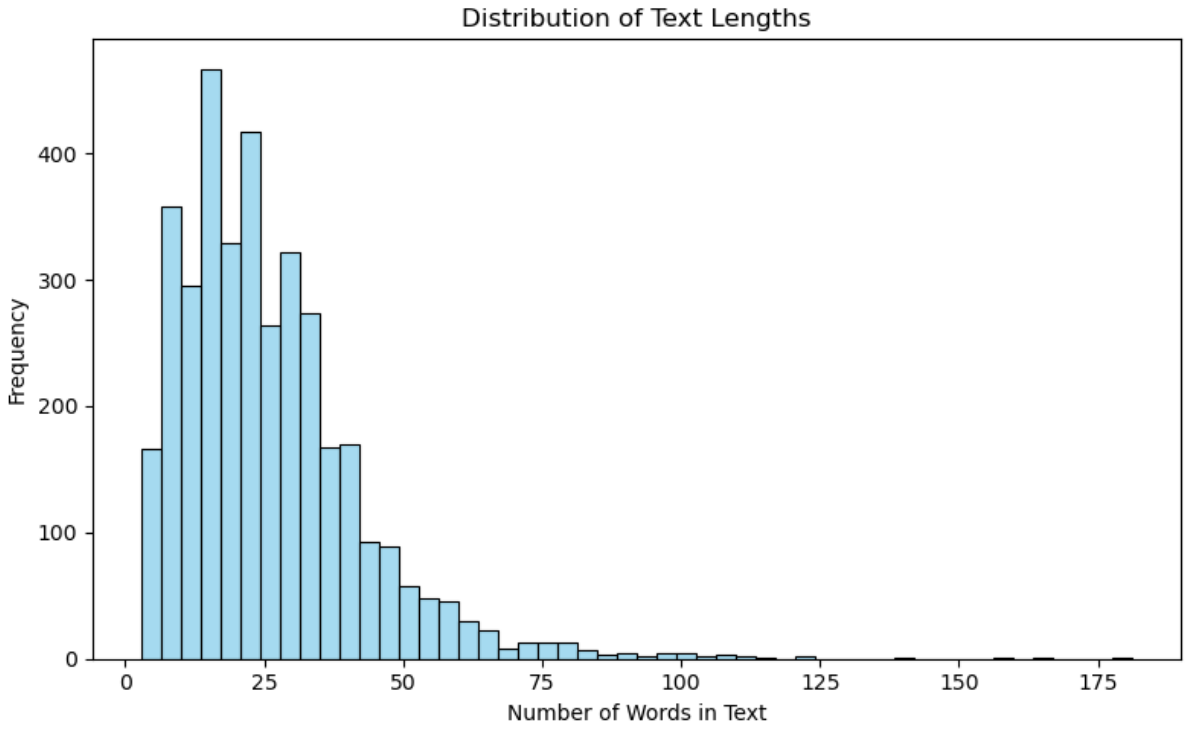

Figure 2: Distribution of Text Lengths Histogram

The histogram seen in Figure 2 shows that most texts fall within 10 to 40 words, with a median of 23 words. A few outliers beyond 150 words are present, but the majority are short texts. These results support the desire to use padded sequences of about 40 to 50 tokens for efficient modelling (OpenAI, 2025).

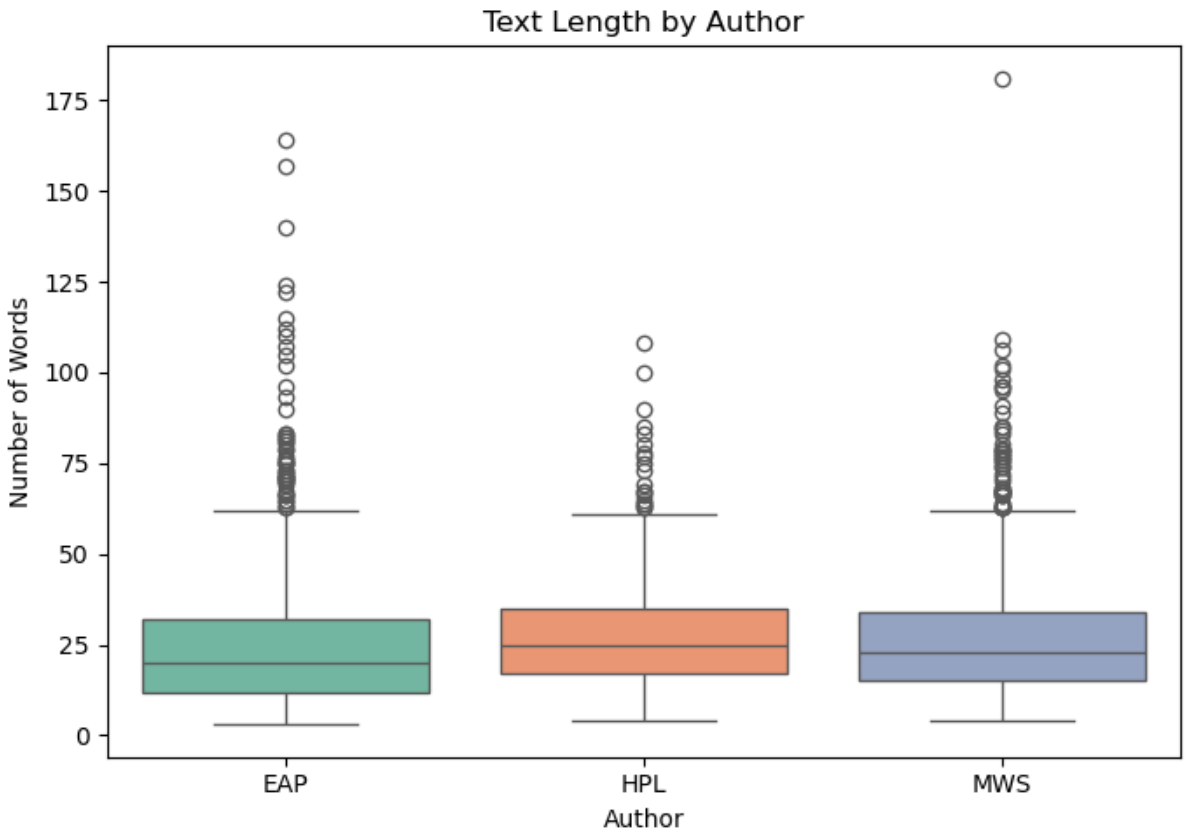

Figure 3: Text Length by Author Box Plot

Figure 3 shows the box plots shows the distribution of text lengths across the three authors (EAP, HPL, and MWS). From Figure 3, we can say that text lengths across authors are similar. However, author MWS has a little longer upper range passages as compared to EAP and HPL. This can be considered good for the model in capturing stylistic differences (OpenAI, 2025).

Modelling Process

The modelling process for this author detection task followed a structured pipeline of data preprocessing, feature engineering, and classification model development using a Long Short-Term Memory (LSTM) recurrent neural network. Each stage of this process was designed to ensure data quality, capture stylistic features in text, and produce results that could be evaluated with robust performance metrics.

Feature Engineering

This project worked with raw text excerpts instead of structured numeric variables. These excerpts were cleaned and tokenised, and then converted into sequences of integers corresponding to a fixed vocabulary. These steps were essential to make these excerpts usable for modelling. The key steps were as follows:

- Text preprocessing: All text within the excerpts was converted to lowercase, and punctuation and special symbols were removed to reduce noise (Muller & Guido, 2016).
- Tokenisation: A vocabulary of 30,000 most frequent words was constructed, then each word was mapped to an index, allowing sequences of integers to represent each excerpt (OpenAI, 2025).
- Sequence padding: Given the varying length of text excerpts, all sequences were padded or truncated to a maximum of 40 words. This provides uniform dimensions for the input into the LSTM (Muller & Guido, 2016).
- Target labels: The author column was encoded into numeric classes, which formed the target variable for the multi-class classification (OpenAI, 2025).

These essential steps gave dense, sequential input data suitable for training an LSTM model (OpenAI, 2025). A model which is ideal for capturing stylistic and contextual dependencies in natural language (Muller & Guido, 2016).

Model Training

The dataset was then split into training, validation, and test sets, following a 70/15/15 split, respectively. This was done to ensure fair evaluation of generalisation (Muller & Guido, 2016). The LSTM architecture included an embedding layer to map word indices to dense vector representations, a dropout layer to reduce overfitting by randomly deactivating units, an LSTM layer for capturing contextual writing patterns across tokens, and a dense softmax output layer to predict the probability of each of the three authors (Muller & Guido, 2016; OpenAI, 2025).

Both the Adam optimiser, which has a learning rate of 0.001, and categorical cross-entropy loss were used to compile the model (OpenAI, 2025). Additionally, early stopping was enabled to halt training if the validation loss stopped improving, which avoids overfitting and unnecessary epochs (Muller & Guido, 2016).

After the model evaluation process, which involved assessing the model based on the following performance metrics and visuals: accuracy, precision, recall, F1-score, macro ROC-AUC, confusion matrix, and learning curves. The model was retrained with adjusted hyperparameters to test the robustness and generalisability. The adjusted hyperparameters included a higher dropout rate and L2 penalties, and adapted learning rate scheduling. These modifications allowed evaluation of the impacted model performance, particularly in the reduction of overfitting and confusion between authors.

Model Evaluation

Now that the model has been trained, it is essential to interpret and evaluate its performance using standard classification metrics. The evaluation draws on Figures 4, 5, and 6, below. Being the Confusion Matrix, the Learning Curve: Loss, and the Learning Curve: Accuracy, respectively. This combination offers a comprehensive view of the model's effectiveness and training dynamics (Muller & Guido, 2016).

Model Performance

The Long Short-Term Memory (LSTM) neural network was trained on the Spooky Author Identification dataset and evaluated on the held-out test set. The performance was assessed using the following metrics: accuracy, precision, recall, F1-score, ROC-AUC, confusion matrix (Figure 4), and learning curves (Figures 5 and 6).

- Accuracy: 77.36% - Shows an overall measure of correctness, but can mark per-class disparities in multi-class settings like this one (Muller & Guido, 2016). This value shows the model correctly predicted the author of about 77% of texts.
- Precision: 78.20%, Recall: 77.36%, F1-score: 77.50% - These metrics reveal how well the model performs for each author. Precision shows the correctness of predicted labels, recall shows how many true labels were captured, and F1-score balances precision and recall (Muller & Guido, 2016). These results show a balance across classes, although precision is slightly higher than recall, meaning fewer false positives than false negatives (OpenAI, 2025).
- ROC-AUC Score: 0.919 – This metric is threshold independent and ensures equal weighting across authors, reducing the risks of bias towards a majority class (Muller & Guido, 2016). The value of 0.919 shows strong separability between authors' styles (OpenAI, 2025).

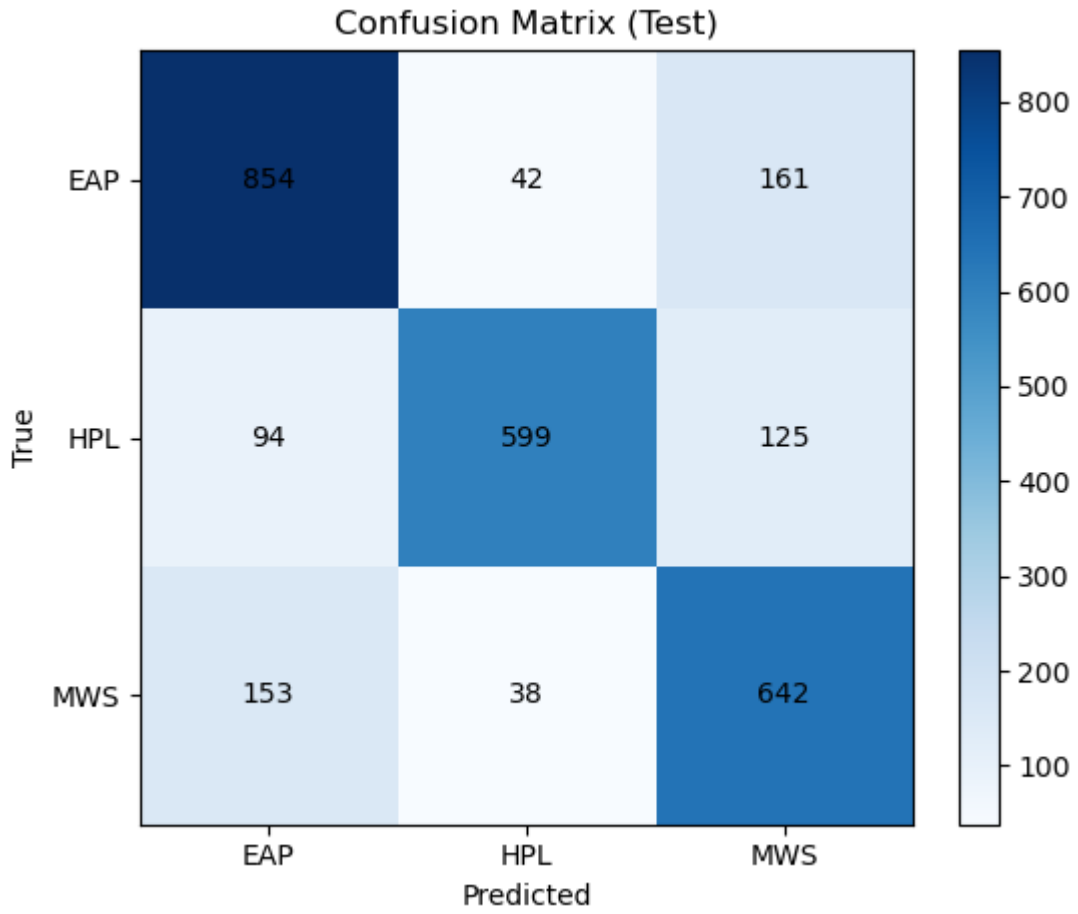

Figure 4: Confusion Matrix Heatmap – Baseline Model

The confusion matrix seen in Figure 4, EAP (Edgar Allan Poe), is the most reliably recognised, with 854 correct predictions. Misclassifications were primarily with MWS (Mary Shelley) at 161 texts. As for HPL (H.P. Lovecraft), there was some confusion with EAP at 94 texts, but overall had strong recognition with 599 correct predictions. Finally, MWS achieved 642 correct predictions; however, it overlapped with both EAP and HPL, given the stylistic similarities across Gothic literature (OpenAI, 2025). Overall, most texts were correctly attributed to their authors.

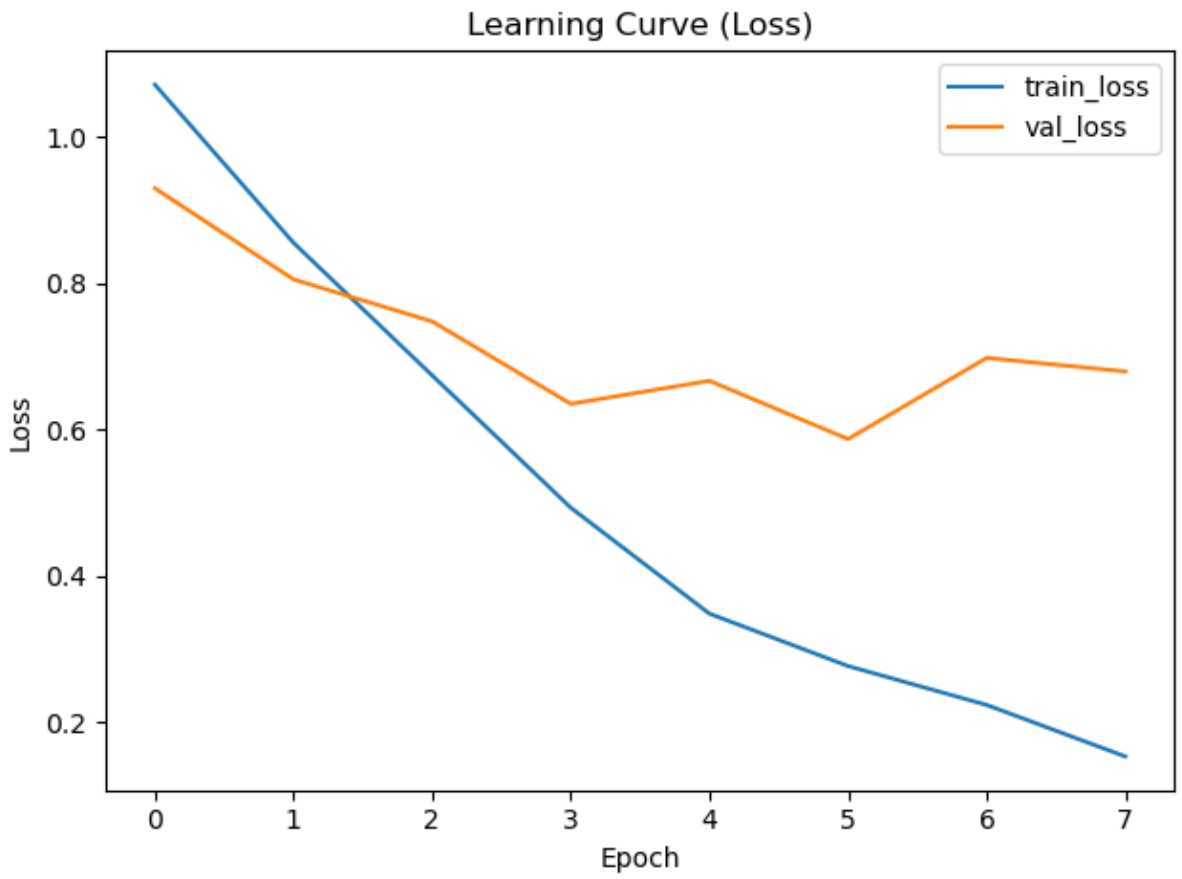

Figure 5: Learning Curve: Loss – Baseline Model

The learning curve seen in Figure 5 reveals that training loss decreased steadily, whilst validation loss flattened after epoch 3-4 and fluctuated slightly (OpenAI, 2025). This suggests modest overfitting (Muller & Guido, 2016).

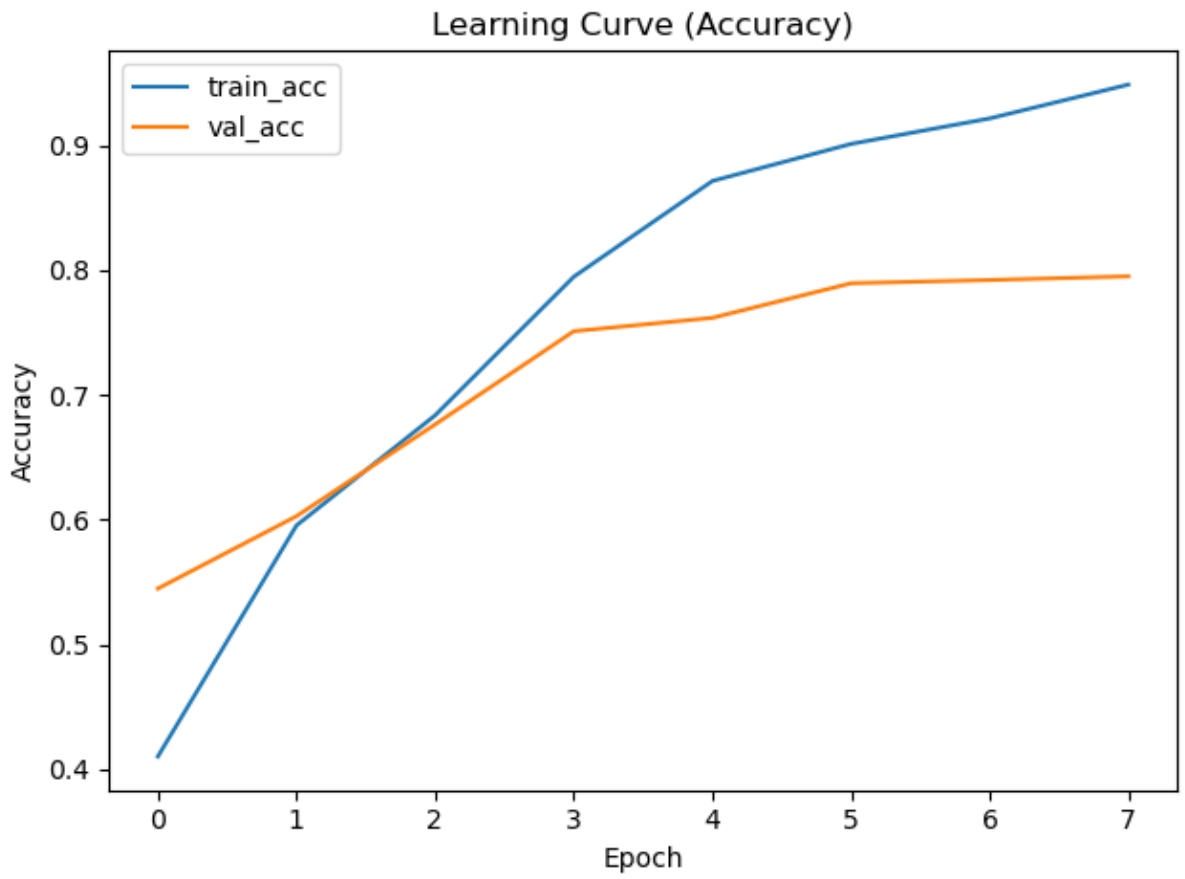

Figure 6: Learning Curve: Accuracy – Baseline Model

Figure 6 above shows that training accuracy rose above 94%, although validation accuracy plateaued around 79%. This gap between curves suggests that the model potentially learned author-specific quirks in training data (OpenAI, 2025). Both Figures 5 and 6 show the validation performance stabilised, meaning the model captured broadly generalisable stylistic patterns (Muller & Guido, 2016; OpenAI, 2025).

Evaluation Summary

The LSTM model showed strong performance in distinguishing between the three Gothic authors. An ROC-AUC value of 0.919 shows the model can separate classes beyond raw accuracy (OpenAI, 2025; Muller & Guido, 2016). Although mild overfitting is visible in the learning curves (Figures 5 and 6), the model generalises sufficiently well. The challenges of overlapping vocabulary and tone are highlighted in the confusion between HPL and EAP.

Model Retraining

Despite the strong baseline performance with an accuracy of 77.4%, an ROC-AUC score of 0.919, and balanced precision and recall, there was still confusion between MWS and HPL and overfitting, given that the validation accuracy plateaued at 79% and training reached over 94% (Muller & Guido, 2016). A reliance on memorised token sequences is suggested instead of general stylistic cues (OpenAI, 2025). To combat this, the model was retrained with adjusted hyperparameters that included a higher dropout rate and L2 penalties, and adapted learning rate scheduling. These changes aim to reduce misclassification, improve generalisation, and reduce overfitting, whilst maintaining high ROC-AUC (OpenAI, 2025). The same performance metrics were considered for evaluation.

Retrained Model Performance

- Accuracy: 79.73% - This result improved from the baseline model, confirming the retraining boosted robustness.
- Precision: 80.26%, Recall: 79.73%, F1-score: 79.59% - These metrics all improved from the baseline model, showing a balanced performance across all three authors, with precision being strongest for HPL and recall strongest for EAP (OpenAI, 2025).
- ROC-AUC Score: 0.93 – This metric rose to 0.93 from 0.919 (baseline model), showing an excellent ability to separate authorial styles beyond accuracy.

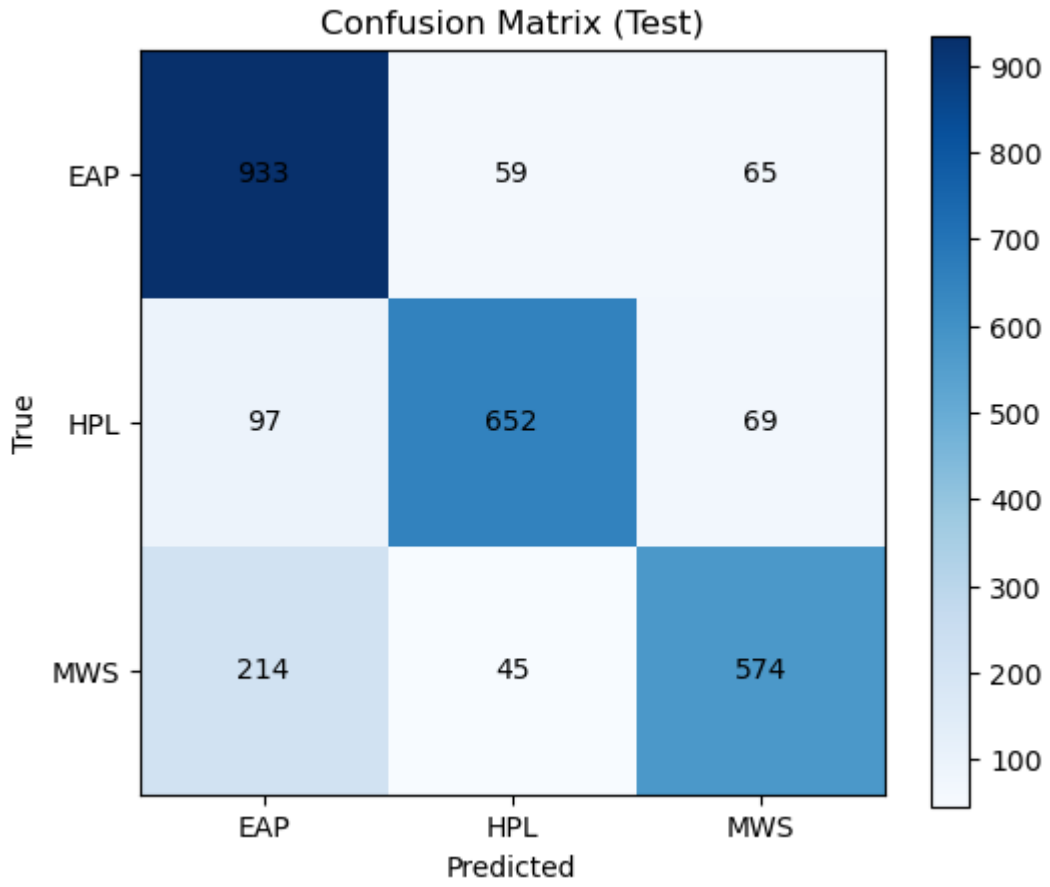

Figure 7: Confusion Matrix Heatmap – Retrained Model

The confusion matrix above reveals EAP had the highest recall at 933 correct predictions, although some overlap remains with MWS. HPL showed a strong balance, misclassifying 97 into EAP and 69 into MWS. Finally, MWS showed some overlap with Gothic vocabulary, with confusion towards EAP at 214 misclassifications.

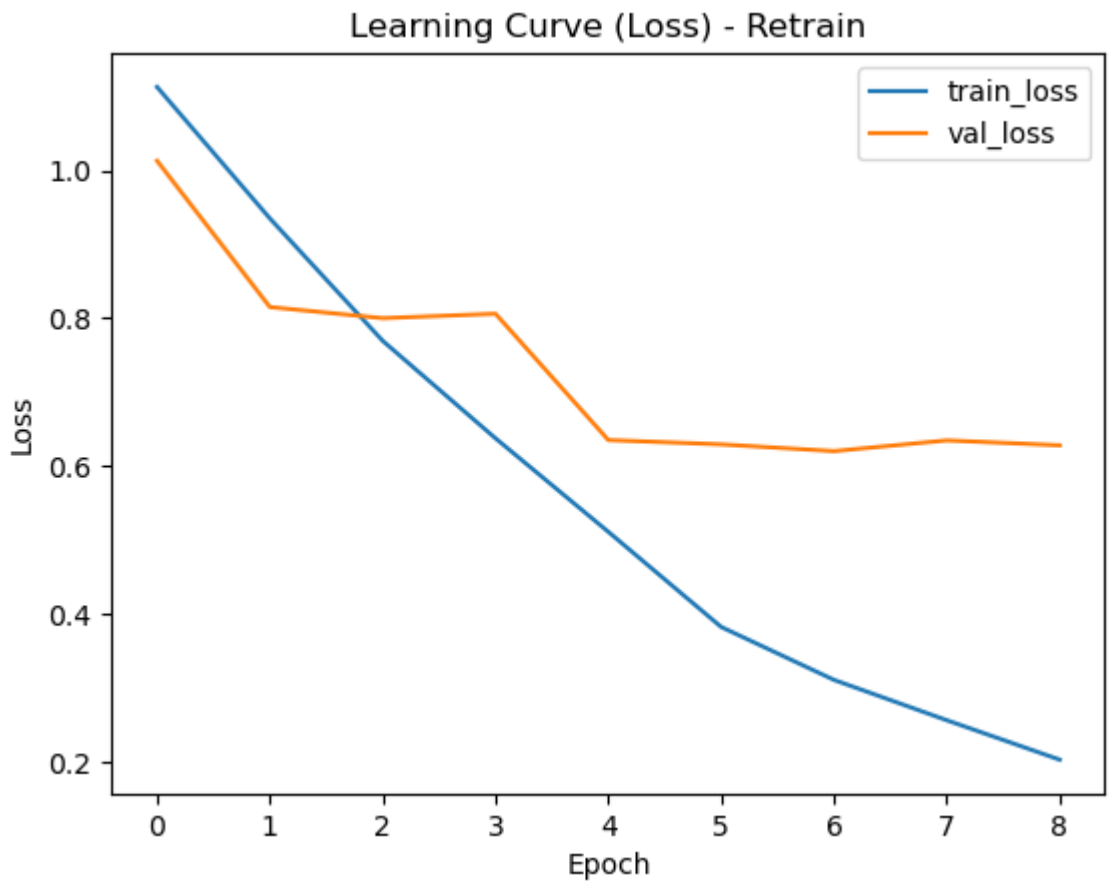

Figure 8: Learning Curve: Loss – Retrained Model

The learning curve for loss above shows validation loss seems to stabilise instead of sharply increasing, showing a reduction in overfitting compared to the baseline model (Muller & Guido, 2016; OpenAI, 2025).

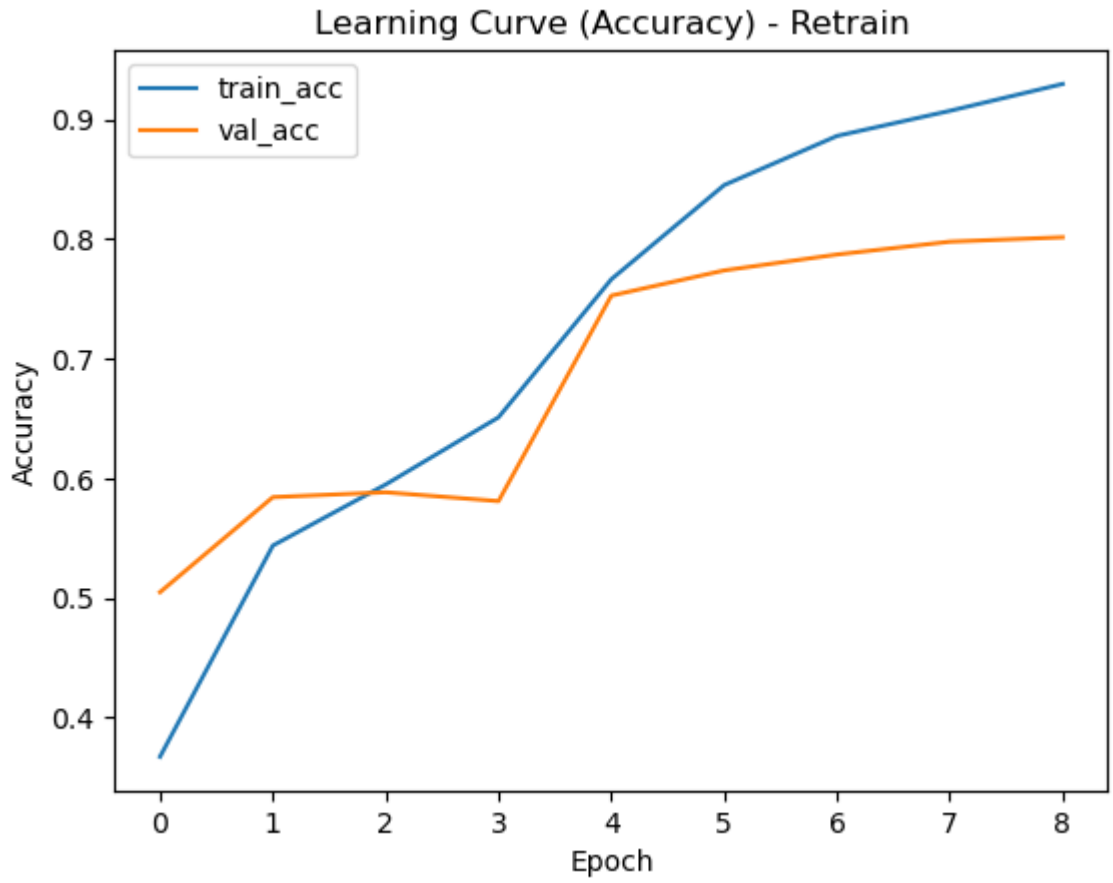

Figure 9: Learning Curve: Accuracy – Retrained Model

Figure 9 above reveals training accuracy scaled over 90% whilst validation accuracy plateaued just under 80%. It can be said that the early stopping and learning rate reduction aided in maintaining generalisability whilst keeping accuracy (OpenAI, 2025).

Retrained Evaluation Summary

In conclusion, the retraining showed the LSTM's ability to generalise stylistic differences between authors and improved model stability. The retrained model is more suitable for deployment given the performance results, reduced overfitting, and slight gains in accuracy and ROC-AUC. For future gains, deeper architectures or data augmentation are appropriate, as there is persistent misclassification between Gothic authors, given the inherent stylistic overlap (OpenAI, 2025).

Conclusion & Recommendations

Recommendations based on findings are as follows:

1. Adopt the Retrained LSTM model: The retrained model showed improved generalisation with an accuracy of about 79.7% and ROC-AUC of 0.93. These results, along with the reduced overfitting, make it the preferred model for deployment over the baseline model.
2. Mitigate Stylistic Overlap with Advanced Techniques: The persistent misclassification shows the challenge of distinguishing authors with overlapping Gothic vocabulary. In the future, deeper architectures like Bi-LSTM or transformers should be applied to capture more nuanced stylistic differences (OpenAI, 2025).
3. Use Data Augmentation to Strengthen Training: To further boost robustness, reduce confusion between authors, and increase stylistic diversity, the addition of synthetic data, like paraphrasing or back translation, would be appropriate (OpenAI, 2025).
4. Continue Evaluating and Tuning: Ongoing monitoring of the model is essential, given that it is sensitive to overfitting. To maintain performance stability, the use of techniques like cross-validation, adaptive learning rate schedules, and embedding dimension tuning should be revisited (OpenAI, 2025; Muller & Guido, 2016).

Conclusion

To summarise, this project effectively showed the use of Long Short-Term Memory (LSTM) recurrent neural networks for multi-class text classification, applied to the Spooky Author Identification dataset. The model chosen for deployment achieved a 79.7% accuracy and a strong macro ROC-AUC of 0.93, confirming its ability to capture distinct authorial styles. The retraining with stronger regularisation improved robustness, reduced overfitting, and delivered more balanced performance across authors. Although the stylistic overlap in Gothic writing aided in the persistent misclassifications, the chosen model offers a reliable foundation for author identification tasks. More so, it highlights opportunities for further performance enhancement using advanced architectures like transformers, as highlighted in the recommendations section above.

Disclosure of AI Use

Sections: Part 1.

Name of the tool used: ChatGPT5.

Purpose behind use: Outlines, summaries, Python code, queries, evaluations, suggestions, and paraphrasing.

Date used: 06/09/2025.

Link to chat: https://chatgpt.com/share/689c605a-7420-8004-8afe-fc6317e663de

References

Kaggle, 2017. *Spooky Author Identification.* [Online]
Available at: https://www.kaggle.com/competitions/spooky-author-identification/overview
[Accessed 07 September 2025].

Muller, A. C. & Guido, S., 2016. *Introduction to Machine Learning with Python.* 1st ed. Sebastopol: O'Reilly Media.

OpenAI, 2025. *Open AI ChatGPT5.* [Online]
Available at: https://chatgpt.com/share/689c605a-7420-8004-8afe-fc6317e663de
[Accessed 07 September 2025].