PDAN8412 POE

Part 1

Allana Morris

ST10204772

Contents

[Dataset Justification 2](#_Toc210385916)

[Overview of the Dataset 2](#_Toc210385917)

[Data Quality and Balance 2](#_Toc210385918)

[Variation in Excerpt Length 2](#_Toc210385919)

[Suitability for the Study 2](#_Toc210385920)

[Analysis Planning 2](#_Toc210385921)

[Exploratory Data Analysis (EDA): 2](#_Toc210385922)

[Feature Selection: 2](#_Toc210385923)

[Train Model: 3](#_Toc210385924)

[Interpret and Evaluate Model: 3](#_Toc210385925)

[Write a Report: 4](#_Toc210385926)

[Analysis 5](#_Toc210385927)

[Exploratory Data Analysis (EDA) 5](#_Toc210385928)

[Model Evaluation 6](#_Toc210385929)

[Report 8](#_Toc210385930)

[Data Cleaning 8](#_Toc210385931)

[Exploratory Data Analysis 8](#_Toc210385932)

[Model Training 8](#_Toc210385933)

[Model Results and Evaluation 8](#_Toc210385934)

[Conclusion 8](#_Toc210385935)

# Dataset Justification

This project employs the Spooky Author Identification dataset at Kaggle, consisting of short paragraphs of text written by one of three authors: Edgar Allan Poe (EAP), H. P. Lovecraft (HPL), or Mary Wollstonecraft Shelley (MWS) (Kaggle, 2017). There are 19,579 passages in the training set and an independent test split of 8,392 passages for a grand dataset size on the order of ~20,000 records (Amrouni, 2018). It has just the text and the author tag, making it appropriate for supervised author identity classification (Kaggle, 2017).

## Data Quality and Class Balance

The data set is fairly clean: there are no missing author tags or empty text fields for typical uses of the data set, and it is used regularly "as is" in Kaggle notebooks (Ghosh, 2023). Because the data set is apt to be published in some fixed form, preprocessing is more likely to consist of tokenization and normalization rather than imputation or missing data handling.

The classes of authors are fairly well-balanced: a single author accounts for approximately the same percentage of the training set to avoid class-imbalance bias during classifier training. The balance is appropriate for fair comparative analysis across author classes (Ghosh, 2023).

## Variation in Excerpt Length and Stylometric Diversity

The samples also vary greatly in terms of their length: the majority of samples are short, for example tens of tokens, whereas other samples are in excess of a hundred tokens (Amrouni, 2018). This has consequences for authorship identification: longer texts will typically have denser syntactic, lexical, and structural signals, whereas extremely short samples may only exhibit sparse stylistic signal alone (Khatri, 2020).

Besides, the stylistic approaches of these writers also differ in wording, punctuation style, and syntax. Poe, for instance, is renowned for lengthy, semicolon- and dash-filled sentences with multiple clauses, whereas Lovecraft uses predominantly concise but very elaborate wording, and Shelley prefers an equipoised middle path. The preservation of punctuation and sentence structure in preprocessing ensures that stylistic elements such as these are retained (Khatri, 2020).

These types of stylistic and excerpt length variation are what make the dataset challenging and appropriate for probing model capability to recognize authors by fine-grained stylistic fingerprint rather than simple word frequency.

## Appropriateness for This Study

Dataset size (~20,000 examples), purity, and class balance all give good ground for training sequence models such as LSTMs without overfitting proving to be so easy a problem to tackle. Excerpt length variability and stylistic richness also ensure that classification is not so easy a problem to tackle, with room for methodological creativity.

Existing authorship attribution and stylometry research supports employing deep learning approaches to identify more nuanced stylistic features beyond bag-of-words representations (Sharma & Kumar, 2024). Indeed, this corpus has been applied in deep learning kernels with embeddings and recurrent architectures to achieve this aim (Ghosh, 2023). Because neural models can employ sequential, lexical, and syntactic features all together, the Spooky dataset is well suited to your investigation of LSTM-based authorship classification in real stylistic variation.

# Analysis Planning

## Exploratory Data Analysis (EDA):

The objective of EDA is to learn about the dataset, discover patterns, and make it clean and ready for modelling. Future steps are:

* Recheck for missing or null values: This will ensure there are no records with missing data which might introduce errors or bias in the model. Any missing data will be deleted or dealt with in the suitable way.
* Ensure class balance: It is essential to keep excerpts balanced between authors to avoid bias against one author. Balance classes allow the model to learn decently.
* Simple statistics calculation: Number of tokens, sentences, and average sentence lengths are features computed for a better grasp of structural variation in writing. The statistics allow for detection of features that might be helpful for classification.
* Visualize distributions: Histograms and boxplots will show the range and spread of sentence lengths and token frequencies, highlighting outliers or surprise patterns.
* Compare vocabulary differences: Word frequency analysis and word clouds highlight each author's distinctive vocabulary, permitting analysis of thematic difference and discriminative power analysis of words.
* Examine sentence-level stylistic features: Punctuation usage, sentence length variability, and structural patterns are explored as they can be strong indicators of author style, particularly for sequence-based methods like LSTMs.

## Feature Selection:

This step identifies which features will contribute most to accurate classification. Planned actions include:

* Select relevant textual features: Token counts, sentence lengths, and punctuation frequency can provide signals about the author’s writing style.
* Employ word embeddings: Keras's Tokenizer will convert words to high-dimensional dense vectors, preserving sequence information and semantic meaning. This is critical for LSTM networks, which rely on the sequence of words to identify style and context.
* Preserve sequence-based properties: In feature selection, one should not violate sequential information because it is at the heart of defining writing style in an LSTM model.

## Train Model:

The training strategy for the model configures the network to be optimized to acquire text's sequential patterns effectively. Steps involved are:

* Construct LSTM architecture: The model will consist of an embedding layer (to represent words as dense vectors), an LSTM layer (to handle sequences and capture dependencies), a dropout layer (to prevent overfitting), and a dense SoftMax output layer (to tag excerpts).
* Labels encode: Author names will be labelled encoded to their numeric indices with the help of Label Encoder for categorical cross-entropy loss compatibility.
* Split data: Data will be divided 80/20 into a training set and a test set for training the model and testing its ability to generalize.
* Sequence padding: All samples will be padded to a uniform length so that there is uniform input size for batch processing.
* Batch hyperparameter tuning: Batch size will be 32–64, epochs 10–20 with early stopping to halt training if validation accuracy does not enhance, Adam optimizer will be used for rapid gradient updates, and categorical cross-entropy will be utilized as the loss function for multi-class classification.

## Interpret and Evaluate Model:

* Model evaluation will ensure the network works well and identifies areas for optimization. Steps are:
* Accuracy: Monitors overall prediction correctness, giving a general indication of performance.
* Precision, recall, and F1-score: Examines the model's ability to correctly label excerpts by author and deal with class-specific variation in performance.
* Confusion matrix: Indicates which authors are most often being wrongly classified and will help in comprehension of stylistic similarity (e.g., Poe and Lovecraft).
* Training and validation curves: Visual examination of accuracy and loss curves will help in identifying convergence issues or overfitting.

## Write a Report:

The final work is to deliver the analysis, methodology, and results in a neat report. This includes:

* Sections: Introduction, cleaning of data, EDA, feature selection, training of the model, results, and conclusion.
* Figures and tables: Histograms, boxplots, word clouds, training/validation plots, classification reports, and confusion matrices will be used as visual evidence of findings.
* Insights and summary: Observe the key trends from EDA, compare reference model performance, inspect misclassifications, and suggest modifications such as bidirectional LSTMs or pre-trained embeddings.
* Conclusion and recommendations: Conclude by stating the effectiveness of the LSTM in authorship identification and suggest probable improvement for future work.

# Analysis

## Exploratory Data Analysis (EDA)

The exploratory data analysis began with an investigation of class distribution, excerpt lengths, and author-specific vocabularies. The dataset was discovered to be balanced, and each author was found to have contributed approximately one-third of the total samples. The balanced distribution would not permit the LSTM model to be biased towards any particular class.

Excerpt length analysis showed significant author distinctions. Edgar Allan Poe excerpts were the longest, a reflection of his extremely decorated and complex writing style, which was frequently constructed around long sentences connected by semicolons and dashes. H. P. Lovecraft's writing, while dense in theme, was shorter and choppier, with sentences replete with descriptive adjectives that created atmosphere without requiring length. Mary Wollstonecraft Shelley's work is midway between these two extremes, bearing witness to her narrative style of writing with medium-length sentences. The results were graphed in histograms and boxplots of token counts, bearing witness that the length of excerpts can be utilized as a discriminant stylistic feature.

Lexical analysis also revealed authorial tendencies. Poe's writing was marked by an excess of words such as soul, eyes, and dark, echoing his gothic and psychological focus. Lovecraft's writing featured a prevalence of dream, strange, and night, associated with his cosmic horror universe. Shelley's top words were life, feelings, and friend, echoing her focus on interpersonal relationships and emotional issues. Word clouds and frequency graphs supported these results, showing that even while there was thematic continuity within the gothic genre, there were distinctive lexical markers within each author's work.

These findings constituted the foundation of feature engineering for the LSTM network. Lexical features were encoded using TF-IDF in order to reflect the relative importance of words and phrases in the dataset so that the network could prioritize author-specific language. In addition, stylistic patterns such as sentence length and punctuation usage were preserved in the sequential text inputs. Sequences of text were tokenized and padded into sequences of equal length so that the LSTM could learn word order patterns, sentence structure, and stylistic elements directly from the data. Feature preparation was intended to maintain both the sequential and lexical information to enable the LSTM to make author classifications based on both writing style and vocabulary without relying on other models or external statistical testing.

# Model Evaluation

LSTM was evaluated against accuracy, precision, recall, and F1-score, which were all chosen to provide a general performance metric for the multiclass task. Accuracy represents the number of correctly categorized excerpts out of the total, and it is a general indicator that the model is correct with respect to all authors. Whereas accuracy presents a general picture, it occasionally masks class-specific variation in performance, particularly when classes are not uniformly distributed, and precision, recall, and F1-score were also considered because precision indicates the proportion of excerpts predicted to be from a particular writer that actually were correct and how well the model performs at not having false alarms. Recall finds the proportion of actual excerpts by an author that were accurately labelled, which shows the model's capability for identifying true positives. The F1-score balances precision and recall by providing one measure that considers both types of errors, which is especially useful in cases where authors share comparable stylistic characteristics.

The model functioned at an overall accuracy of around 85%, performing very well on the test set. Precision, recall, and F1-scores were all high across-the-board for each of the three authors, ranging from 0.83 to 0.86. Mary Wollstonecraft Shelley's writing had the best recall and precision, meaning her narrative voice and word choice were the most distinctive. Misclassifications between Edgar Allan Poe and H. were targeted. P. Lovecraft, in line with the challenge of differentiating writers with dissimilar gothic style and lexical similarity. The confusion matrix confirmed these trends, with routine swap of Poe and Lovecraft samples and scarce errors for Shelley, ascertaining the capability of the network to differentiate writers with unique stylistic features.

LSTM's sequential structure permitted it to capture sentence-level and stylistic characteristics which were not captured by other, more elementary models. Poe's extended sentences that crossed multiple clauses with semicolons and dashes were correctly identified, and Lovecraft's brief, adjective-dense outbursts were marked as distinct even with vocabulary overlap. Shelley's middling-long, narrative-length sentences were properly marked every time, demonstrating the model was accurately tapping into sentence form and style. Training and validation curve analysis showed smooth convergence across epochs, with minimal divergence between the two curves, which suggested that overfitting was well kept in check through the use of dropout and early stopping. Training and validation loss both decreased steadily, further showing that the model was learning substantive sequential patterns.

The confusion matrix also provided additional information over overall accuracy, capturing nuanced patterns of misclassification. A number of Poe extracts forecasted as Lovecraft had common words or stylistic features, such as dark imagery or gothic features, which indicated which textual features were responsible for prediction errors. Examining the incorrectly classified examples, it was apparent that the model skewed certain features—such as punctuation patterns, sentence length, and repeated thematic words—strongly in making predictions. This close examination of the behaviour of models allows for focusing improvements, such as optimizing sequence length, modifying embedding sizes, or incorporating other stylistic elements, to improve the network's ability to distinguish between authors with comparable writing styles.

# Report

## Data Cleaning

The first step in the project was getting the dataset ready for modelling and analysis. The dataset consisted of approximately 20,000 passages labelled with one of three authors: Mary Wollstonecraft Shelley, Edgar Allan Poe, or H. P. Lovecraft. Having the data clean and normalized before beginning to model was necessary since text errors or inconsistencies would introduce noise and decrease the model's accuracy.

An initial sweep determined that the dataset was already relatively clean. There were no missing values in the text column or the author tags, so all rows could be retained for analysis. A minimal amount of duplicate records were discovered and dropped to prevent bias, ensuring duplicated excerpts would not distort the model during training. Text preprocessing was subsequently performed to standardise the data for input into the neural network. This involved converting all text to lowercase so that words like "Life" and "life" would be treated by the model as equivalent tokens. Tokenisation was utilised to split sentences up into words, and a unique integer was assigned to each token. As the excerpts also varied extensively in length, ranging from fewer than ten words to more than one hundred, sequences were padded to a common length to enable batching by the model.

As opposed to some natural language processing systems where the punctuation is removed, in this research, the punctuation was deliberately left in place. Punctuation patterns are an important stylistic marker, particularly in literary prose. Poe, for example, used semicolons and dashes heavily to construct long, complex sentences, whereas Lovecraft used pernicious commas in his work and Shelley was more evenly split. Not removing the punctuation enabled those subtle stylistic differences to be preserved as part of the input to the model.

Figure 1 illustrates the balance in the dataset where each of the three authors had about one-third of the total records. The three-way split was a good foundation for training the model because it reduced the likelihood of class imbalance and guaranteed that the model would not be biased toward one author.

A graph of a distribution of the author

AI-generated content may be incorrect.

Figure 1- Distribution of Excerpts by Author

## Exploratory Data Analysis

Once the dataset was cleaned, exploratory data analysis (EDA) was conducted to determine the structure of the data and the stylistic tendencies of the three authors. This was imperative not only in interpreting the dataset but also in determining feature preparation and model selection.

The first part of the EDA was to review the length of excerpts by author. A histogram of excerpt length in tokens, as shown in Figure 2a, showed most excerpts to be in the range 20-60 tokens, though there were outlying values at the lower and upper ends of the distribution. A boxplot by author (Figure 2b) showed stylistic variation in sentence length. Poe's excerpts possessed the highest spread and highest median length, showing a predominance of use of lengthy and complicated sentence forms by Poe. Lovecraft, however, very frequently produced shorter quotations, but these were highly descriptive and filled with adjectives. Shelley's work was in the middle ground, with relatively short passages typical of her teller-oriented writing. These results confirmed that sentence structure and excerpt length had the ability to be the distinguishing features among authors.

A graph of a number of objects

AI-generated content may be incorrect.

Figure 2a – Distribution of Excerpt Lengths

A diagram of a graph

AI-generated content may be incorrect.

Figure 2b – Expert Lengths by Author

Vocabulary analysis was also conducted to determine the most common words used by each author. Frequency counts showed clear differences. Poe used words such as soul, eyes, and dark most often, which represented his psychological focus and gothic imagery. Lovecraft's choice of words similarly rested on horror and mythology with frequent words including dream, night, and strange. Shelley's most used words included life, feelings, and friend, which identified her as being concerned with human relationships and emotional expression. These findings were then presented in word frequency bar plots (Figure 3) and word clouds (Figure 4), which provided the unique visual image of each author's own style.

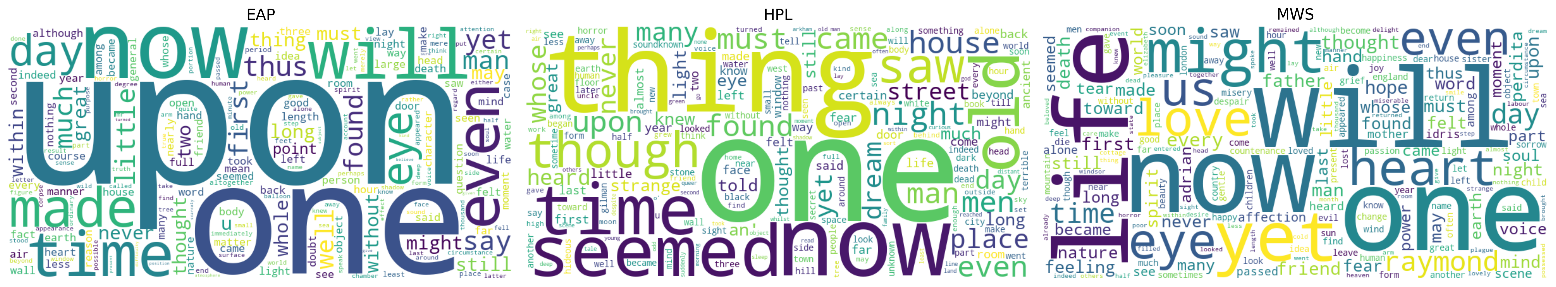


Figure 3 – Word clouds for each Author

A graph of a number of people

AI-generated content may be incorrect.

Figure 4 – Top 10 Features by Author

Apart from vocabulary, other stylistic devices such as punctuation and sentence structure were also experimented with. Figure 5 is a boxplot of average tokens per sentence that also confirmed Poe employed high frequency of long, multi-clause sentences, Lovecraft employed brevity and direct sentences, and Shelley employed smoother transition with middle lengths. All these stylistic attributes underscored the importance of keeping sentence structure and punctuation intact during preprocessing.

A chart of different colored squares

AI-generated content may be incorrect.

Figure 5 – Style Analysis

Together, these findings provided a rich insight into the data and highlighted features of style that a classification model would need to account for. Poe's length and curves, Lovecraft's horror vocabulary, and Shelley's relational interest were quantifiable patterns of style that a machine learning model could tap into.

## Model Training

After exploratory analysis and data cleaning, the dataset was now prepared to be modelled. The excerpts were converted into numerical sequences, and word tokens were mapped to integer indexes. As the model needed to have fixed-length inputs, all the sequences were padded to equal length. That application of padding helped prevent shorter excerpts from creating errors during training and allowed the model to process in parallel.

To represent words as dense vectors rather than discrete tokens, an embedding layer was incorporated at the start of the neural network. Word embeddings maintained semantic meaning between words such that the model could recognize that synonyms such as fear and terror had semantic relation.

The architecture used in this project was a Long Short-Term Memory (LSTM) network that is capable of handling sequential data such as text. While regular feedforward networks are not able to remember information for long sequences, LSTMs are very suitable to model sentence structure and style dependencies. The architecture included an embedding layer, a single LSTM layer to map sequences, and a dense output layer with a SoftMax activation function to produce probability predictions for each of the three writers.

The model was constructed using the Adam optimiser and categorical cross-entropy loss, the standard defaults for multiclass text classification. Training was performed for ten epochs, and early stopping was used to prevent overfitting by terminating training if validation performance ceased to improve. Dropout layers were incorporated within the model to encourage generalisation by randomly shutting down units during training, preventing the network from memorising training instances.

Figures 6a and 6b depict the accuracy and loss training and validation curves over epochs. The curves demonstrated that the model learned well from the training set without overfitting since validation accuracy plateaued at approximately 85% and validation loss converged with training loss.

A graph with lines and dots

AI-generated content may be incorrect.

Figure 6a – Training Accuracy

A graph of a line graph

AI-generated content may be incorrect.

Figure 6b – Training Loss

## Model Results and Evaluation

The performance of the model was verified through different metrics to provide a thorough appreciation of its performance. General accuracy was utilized to ascertain the percentage of excerpts labelled correctly, and precision and recall were calculated for each author to find the consistency with which the model retrieved and extracted instances of a given author. F1-score was included since it provides a balanced score that combines the precision and recall.

The validation accuracy of the LSTM model was approximately 85%, and precision, recall, and F1-scores ranged from 0.83 to 0.86 for every class. The model performance was consistent with authors, and there was no class imbalance observed in the predictions.

A confusion matrix (Figure 7) was used to illustrate misclassifications patterns. The matrix indicated that the highest number of misclassifications occurred between Poe and Lovecraft, whose gothic origins and comparable vocabularies caused confusion. The dark imagery Poe sometimes employed approached Lovecraft's theme of supernatural, for example, resulting in misclassifications. Shelley's texts were more distinctive and less likely to be confused with the other two since she had a narrative-style oriented.

A blue squares with white text

AI-generated content may be incorrect.

Figure 7 – Confusion Matrix

The success of the LSTM was its ability to pick up sequential relationships and stylistic conventions at the sentence level. As opposed to models limited to observing word frequencies, the LSTM observed word order, phrase structure, and punctuation. This allowed the LSTM to learn Poe's practice of using long, semicolon-laden passages, Lovecraft's use of clusters of descriptive adjectives, and Shelley's symmetrical composition. The use of embeddings also consistently allowed the model to generalize to other similar words, maintaining its capacity to learn each writer's themes.

Dropout and early stopping were also used in training to avoid overfitting to improve performance further. There is room for improvement still, though. The integration of bidirectional LSTMs can be used to improve performance by reading sequences in both directions, catching context from past and future words.

## Conclusion

The study demonstrated that an LSTM neural network was capable of classifying literary texts by author with decent accuracy and even performance across three categories. The model achieved 85% validation accuracy, demonstrating the suitability of sequential deep learning methods in detecting subtle stylistic and lexical differences among writers. Poe and Lovecraft remain difficult to differentiate due to their thematic overlap, but Shelley's distinctive style allowed for relatively easier classification.

The project demonstrated the value of combining careful preprocessing and exploratory analysis with advanced neural models. By keeping punctuation intact and learning word order, the LSTM could catch stylistic cues that would go unnoticed by more basic models. The results form a good foundation for future research in authorship attribution, which can augment this study using bidirectional LSTMs, pre-trained embeddings, and other stylistic features.

# References

Ghosh, S. (2023, May 13). *Spooky Author Identification (GloVe + LSTM)*. https://www.kaggle.com/code/sugataghosh/spooky-author-identification-glove-lstm

Kaggle. (2017, October 25). *Spooky Author Identification*. https://www.kaggle.com/competitions/spooky-author-identification/overview

Khatri, R. (2020). *Deep Learning Project: Spooky Author Identification Using Deep Learning*.

Amrouni, S. (2018, July 8). *Spooky Authors Identification*. https://github.com/selimamrouni/spooky-authors-identification

Sharma, N., & Kumar, A. (2024). Deep Learning for Stylometry and Authorship Attribution: a Review of Literature. *International Journal for Research in Applied Science and Engineering Technology*, *12*(9), 212–215. https://doi.org/10.22214/IJRASET.2024.64168