DEGREE PROJECT

Stylometry: Quantifying Classic Literature For Authorship Attribution

A Machine Learning Approach

Donato SCARANO
Jacob Yousif

Master Programme in Data Science
2024

Luleå University of Technology
Department of Computer Science, Electrical and Space Engineering

LULEÅ UNIVERSITY OF TECHNOLOGY

[This page intentionally left blank]

Contents

List of Figures

List of Figures

Abstract

Classic literature is rich, be it linguistically, historically, or culturally, making it valuable for further studies. Accordingly, this project chose a set of 48 classic books to conduct a stylometry analysis on the defined set of books, adopting an approach used by a related work to divide the books into segments, quantify the resulting segments, and analyze the books using the quantified values to understand the books' linguistic attributes. Apart from the latter, this project conducted different classification tasks for other objectives. In one respect, the study used the quantified values of the segments of the books for classification tasks using advanced models like *LightGBM* and *TabNet* to assess the application of this approach in authorship attribution. From another perspective, the study utilized a *State-Of-The-Art* model, namely, *RoBERTa* for classification tasks using the segmented texts of the books instead to evaluate the model's performance in authorship attribution. The results uncovered the books' characteristics to a reasonable degree. Regarding the authorship attribution tasks, the results suggest that segmenting and quantifying text using stylometry analysis and supervised learning algorithms is practical for authorship attribution tasks. However, this approach may require further improvements to yield better performance. Lastly, *RoBERTa* demonstrated high performance in authorship attribution tasks.

Keywords: *Authorship Attribution, Classic Literature Analysis, Clustering, Computational Linguistics, Data Science, Deep Learning, Feature Engineering, Feature Extraction, K-Means, Light Gradient-Boosting Machine (LightGBM), Machine Learning, Multiclass Classification, Natural Language Processing (NLP), Robustly Optimized BERT Pretraining Approach (RoBERTa), Stylometry Analysis, Tabular Attention-based Neural Network (TabNet), Text Mining, t-distributed Stochastic Neighbor Embedding (t-SNE), Transformer Models.*

Preface

For the exceptional guidance, unwavering support, and kindness, it is essential to express gratefulness to Dr. Oluwatosin ADEWUMI, who supported and guided the entire process of this study. His unconditional support and profound knowledge were instrumental in producing this work.

1 Introduction

Classic literature is still viable in modern times, and in different parts of the world, students study classical literature in schools [1][2]. Whether schools and universities teach classical literature for sociological, historical, cultural, or linguistic purposes, prominent classical literature still contains unique linguistic styles and characteristics that make this type of work gain significance to be part of educational curriculum [3][4]. From a linguistic perspective, classical literature can be a valuable resource for advanced text analysis to obtain information, including but not limited to text structure, vocabulary richness, lexical diversity, cultural contexts, and semantics to use such properties to identify authors, improve writing styles by discovering the characteristics of an adequate language as a tool for human expression and communication [5][4].

In recent years, Natural Language Processing *(NLP)* techniques evolved at a fast speed where they offer a wide range of powerful computational abilities, namely Large Language Models *(LLMs)* that adopt the *Transformer* Architecture [6]. The *LLMs* are capable of processing and analyzing large corpora in-depth to perform various linguistic tasks, some of which are text generation, question answering, text translation, sentiment analysis, and feature extraction [7][8]. Consequently, utilizing *LLMs* on classical literature to conduct stylometry analysis may yield better results not realized by previous studies that used the traditional *NLP* techniques. Although *LLMs* offer robust capabilities for conducting linguistics tasks, they are generally intensive resource consumers and require large amounts of data and lengthy training times [9].

Authorship attribution and stylometry are the most promising data science applications of literary analysis. Traditionally, authorship attribution was based on subjective factors such as handwriting analysis or discovering information about an author's personal life. Although this method could be better, it can lead to errors and bias. *AI* can analyze large amounts of text to identify each author's patterns and unique linguistic attributes. Stylometric analysis can detect forgeries and misattributions [10]. In 2018, a study developed a system to identify writing styles for document classifications using stylometry analysis and unsupervised learning. This project intends to apply a similar approach to the latter study to analyze the linguistic features of classic literature and to perform authorship classification using modern supervised machine learning algorithms [11]. Furthermore, this study will also utilize one of the *State-Of-The-Art* models, namely, *RoBERTa*, to evaluate its performance in authorship classification tasks.

1.1 Background

In terms of language, classic literature is a valuable resource because it generally maintains rich language, sophisticated structure, and compelling linguistic features, making it attractive to a wide range of readers [3][12]. Exploring classic literature with the current machine learning algorithms may yield results that previous studies may need to realize. Furthermore, examining classic literature would entail conducting a stylometry analysis to look at it from different linguistic perspectives.

In 2018, *H. Elahi* and *H. Muneer* developed a solution that can classify various writing styles within a document through stylometric analysis by dividing the document into partitions and using different stylometric metrics to compute the linguistic attributes of the document and vectorizing the resulting features to group the writing styles according to the similarities using unsupervised machine learning algorithms [11]. Although the existing *Transformer* models, like *LLM*, can process large corpus and natural language efficiently and conduct different linguistics tasks, there are still different aspects to consider,

one of which is computational resources. *Transformer* models can be resource-intensive and require extensive data. Additionally, explainability regarding how they reason in their predictions is unclear, meaning they may not yield satisfactory information regarding the driving factors that led to their decisions [13][14][15]. In most cases, it is essential to understand, especially in a language, its features and structure. Hence, using a different approach where the basis of the prediction is on known metrics may benefit literacy studies, data practitioners, education institutions, or other entities.

Therefore, this study explores three dimensions that will involve adopting the work approach of *H. Elahi* and *H. Muneer* of segmenting text and quantifies the segments using stylometric analysis and applying the same metrics criteria [11]. First, to analyze classic literature using the quantified values of the text segments of the books to understand their linguistic features, aiming to provide novel insight that contributes to a wide range of domains, including educational institutions, to shed light on the properties that make a language an effective expression tool in written forms. Second, to use the quantified values with advanced supervised machine learning algorithms to evaluate the feasibility of this approach in classification, aiming to provide a somewhat better feasible and interpretable approach in authorship attribution tasks. Third, utilize a *Transformer* model, namely, *RoBERTa* in authorship classification, aiming to evaluate its performance in such tasks.

Table 1.1 lists the subject books for this study, and they are obtained from the *Gutenberg Project* website [1].

Table 1.1: Books Information

Book	**Publication Year**	**Author**	**Genre**
Sense and Sensibility	1811	Jane Austen	Romance
Pride And Prejudice	1813	Jane Austen	Romance
Emma	1815	Jane Austen	Comedy
Frankenstein	1818	Mary Shelley	Gothic Horror
Valperga	1823	Mary Shelley	Historical Novel
The Last Man	1826	Mary Shelley	Apocalyptic Fiction
Oliver Twist	1838	Charles Dickens	Social Novel
Bartleby, the Scrivener	1853	Herman Melville	Short Story
Moby-Dick	1851	Herman Melville	Adventure Fiction
The Piazza Tales	1856	Herman Melville	Psychological Fiction
Madame Bovary	1857	Gustave Flaubert	Literary Realism
A Tale of Two Cities	1859	Charles Dickens	Historical Fiction
The Mill on the Floss	1860	George Eliot	Realist Novel
Great Expectations	1861	Charles Dickens	Bildungsroman
Silas Marner	1861	George Eliot	Realist Novel
Les Miserables	1862	Victor Hugo	Historical Fiction
Salammbo	1862	Gustave Flaubert	Historical Fiction
Sentimental Education	1869	Gustave Flaubert	Bildungsroman
War And Peace	1869	Leo Tolstoy	Historical Fiction
Middlemarch	1871	George Eliot	Victorian Novel
Ninety-Three	1874	Victor Hugo	Historical Novel
Anna Karenina	1877	Leo Tolstoy	Realist Novel
The History of a Crime	1877	Victor Hugo	Historical
The Adventures of Tom Sawyer	1876	Mark Twain	Bildungsroman
The Adventures Of Huckleberry Finn	1884	Mark Twain	Picaresque Novel
What Men Live By	1885	Leo Tolstoy	Philosophical Fiction

Continued on next page

[1] https://www.gutenberg.org

Table 1.1 – *Continued from previous page*

Book	**Publication Year**	**Author**	**Genre**
A Connecticut Yankee in King Arthur's Court	1889	Mark Twain	Satire
The Picture Of Dorian Gray	1890	Oscar Wilde	Philosophical Fiction
Lady Windermere's Fan	1892	Oscar Wilde	Comedy
The Golden Age	1895	Kenneth Grahame	Children's Literature
The Importance of Being Earnest	1895	Oscar Wilde	Comedy
Dracula	1897	Bram Stoker	Gothic Horror
Dream Days	1898	Kenneth Grahame	Collection of Stories
Heart Of Darkness	1899	Joseph Conrad	Psychological Fiction
Lord Jim	1900	Joseph Conrad	Adventure Fiction
The Wonderful Wizard Of Oz	1900	L. Frank Baum	Fantasy
White Fang	1906	Jack London	Adventure Fiction
Ozma of Oz	1907	L. Frank Baum	Children's Literature
A Room With A View	1908	E. M. Forster	Romance
The Wind In The Willows	1908	Kenneth Grahame	Children's Fiction
The Call Of The Wild	1903	Jack London	Adventure Fiction
The Jewel of Seven Stars	1903	Bram Stoker	Horror
Nostromo	1904	Joseph Conrad	Political Fiction
The Marvelous Land of Oz	1904	L. Frank Baum	Children's Literature
The Sea-Wolf	1904	Jack London	Psychological Fiction
Howards End	1910	E.M. Forster	Domestic Fiction
The Lair of the White Worm	1911	Bram Stoker	Horror
A Passage to India	1924	E.M. Forster	Historical Fiction

1.2 Motivation

This study aims to analyze the classic literature to uncover insights that may not have been realized previously and contribute novel insights to various domains, including educational institutions. Apart from the latter, the study also aims to examine an approach of a related work [11] in a different manner, which is authorship attribution using novel and advanced supervised machine algorithms aiming to evaluate the feasibility of this approach for this task and provide an alternative and feasible approach that is interpretable and not computationally intensive. The latter would give insights into applying this approach in authorship attribution and the algorithms used for this task. It is important to remember that computational resources, digital library maintenance, and the exponential growth of data and *AI* are pushing the demand for water and energy worldwide [14][16].

Moreover, the study aims to use *RoBERTa* to evaluate its performance in authorship classification tasks, and that would provide information to compare the application of *RoBERTa* with the other approach.

1.3 Objectives

This section describes the objectives of the research questions of this study.

O1	Preparing and preprocessing the subject literature for the study.
O2	Segmenting the preprocessed books into text segments and performing stylometry analysis to obtain their textual properties and produce a tabular dataset containing the text segments and their properties.
O3	Investigating the distribution of the segmented text by analyzing the proportionality and performing statistical computations.
O4	Investigating the textual properties of the subject books and their variations to understand their impact on readability by utilizing unsupervised machine learning algorithms; *K-Means* clustering, *t-distributed Stochastic Neighbor Embedding (t-SNE)* and statistical methods.
O5	Utilizing the textual properties of the text segments to train two supervised machine learning algorithms, i.e., predictive models; Tabular Attention-based Neural Network *(TabNet)* and Light Gradient-Boosting Machine *(LightGBM)*, to predict the authorship of text segments and assess the effectiveness of this approach.
O6	Utilizing a *Transformer* model, *Deep Neural Network*, namely Robustly Optimized BERT Pretraining Approach *(RoBERTa)* to train it with the text segments to predict authorship in multiclass classification tasks.
O7	Performing topic modeling by utilizing *RoBERTa* to tokenize and embed the text segments and use the resulting embeddings to train *Latent Dirichlet allocation (LDA)* to explore this approach. This objective is merely for further exploratory experiments.

Table 1.2: The objectives of the study.

The objectives address the following research questions:

- **RQ1**: What role do textual features play in the readability and complexity of classic literary texts?
  - **RQ1** aims to investigate the quantified textual properties of the subject books by utilizing an unsupervised machine-learning approach. **O2**, and **O4** aim to address this question.
- **RQ2**: How effective is segmenting and quantifying texts into a numeric tabular dataset and using it with predictive models to predict the authorship of text segments in multiclass classification tasks?
  - **RQ2** aims to investigate the feasibility and effectiveness of quantifying text segments by performing stylometry analysis and using them with predictive models. **O2**, **O3**, and **O5** aim to address this question.
- **RQ3**: To what extent can *RoBERTa* effectively identify the author of text segments across multiple authors?
  - **RQ3** aims to investigate the feasibility and the effectiveness of *RoBERTa* to predict authorship in multiclass classification tasks. **O2**, **O3**, and **O6** aim to address this question.
- **RQ4**: How does the size of the training dataset impact *RoBERTa*'s performance in multiclass classification?

– **RQ4** aims to investigate the impact of the training dataset's size on the model's performance. **O3**, and **O6** aim to address this question.

1.4 Monetization

Authorship attribution and stylometry have numerous real-world possibilities that could be considered, for example, in legal, ethical, and even commercial contexts; they can be crucial for plagiarism, copyright infringement, or fraud cases. In the news and political arena, stylometric techniques can identify patterns associated with fake news and distinguish between original and fabricated content. Other domain uses can be content verification using stylometry to verify the authenticity of the content in question. For Brand Voice Analysis: Some companies analyze their writing style to maintain consistency across produced materials and content personalization by understanding users' writing styles; companies can personalize communication to match their clients' preferences [17][18][19][20].

1.5 Digitalization

Digital literary studies transform how documents are stored, manipulated, and distributed. They revolutionize how information is handled, making it more accessible, efficient, and adaptable to an increasingly connected world. Furthermore, they promote the conservation of literary heritage for future generations [17].

1.6 Novel Perspectives

In text analysis, new boundaries and research directions are emerging. Computational literary analysis, which harnesses the power of *NLP*, faces a formidable challenge in the form of the long and intricate sentences found in novels, which often push the limits of syntactic parsers. There is also a new interest in developing computational models of the literary plot by decomposing it into sub-problems [21].

Lexical Structure Studies are also using *NLP* tools to characterize the lexical components of different authors. The advent of *AI* propels literary analysis to uncharted territories, where comprehensive analyses can be conducted on entire movements or historical periods, unveiling patterns and trends that were hitherto undiscoverable. The relevance of using *NLP* analysis on 19th-century literary masterpieces [22] is manifold, spanning from understanding literature to cultural insights, from understanding an author's style to teaching and learning literature. Furthermore, it promotes the preservation of literature for future generations and can provide new perspectives for Literary Criticism. Sustainability can be considered from different perspectives. On the one hand, it promotes resource efficiency, scalability, and reproducibility, but on the other hand, it can create an environmental burden.

1.7 Contribution

The study aims to revisit classic literature to provide nuanced insights into its linguistic features, which previous studies may not have realized, by employing stylometry analysis and applying modern and advanced models. Additionally, this research can contribute to computational literary studies and shed light on fresh perspectives on authorship classification and stylometry analysis.

1.8 Delimitation

In this study, the scope is 16 authors, with three books per author, making the total number of books 48—the subject books for the study are listed in Table 1.1. Section 1.3 outlines the goals and the research questions, and this study's boundaries are to conduct the defined objectives to address the defined research questions. The study will use only the technologies, techniques, models, and tools discussed in Section 1.3 and Section 2. In addition, this study will utilize multiple machine learning algorithms, requiring time and computational resources. Due to the limited computational resources, it may not be applicable to execute post-learning processes; hence, performing a post-learning optimization process is out of the scope of this study.

1.9 Ethics

The study analyzed a set of classic literature using different machine learning algorithms. Still, observing the basic ethical principles during the analysis remains essential, as well as maintaining transparency, inclusivity, and fair use. The main concerns regarding literature are respect for intellectual property, data privacy, transparency, fair usage, and inclusivity. Consequently, the study took multiple considerations into account. The first consideration was using books in the public domain that do not require licenses, consents, or other considerations. The latter implies that the books in question do not contain personal information and are not subject to copyright restrictions, allowing their use for academic and study purposes.

Secondly, the study conducted systematic data preprocessing to ensure that data is accurate, present the books and their authors to provide fairness and prevent unintentional bias in the study analysis. Thirdly, the study outlined the methodology, techniques, steps, and processes conducted during its cycle to allow practitioners to review the work, conduct future work based on it, replicate it, or provide a critique.

Lastly, given the nature of the books, that is, classic literature, implying that they come from different eras and contain historical and cultural contexts, the study aims to conduct stylometry analysis using various techniques to analyze the textual characteristics of the books and the methods in question rather than providing any interpretation of the historical and cultural contexts of the subject books.

2 Methodology

This section presents and describes the study's methodology and explains its validity. The methodology involves utilizing multiple algorithms requiring different statistical and mathematical computations to quantify the qualitative data. Hence, this study used quantitative methods to address the research questions.

2.1 Method

The method of this study comprises various phases, and this section outlines each stage of the process.

2.1.1 Data Preparation and Preprocessing

Figure 2.1 demonstrates the process of data preparation and processing from a high-level perspective.

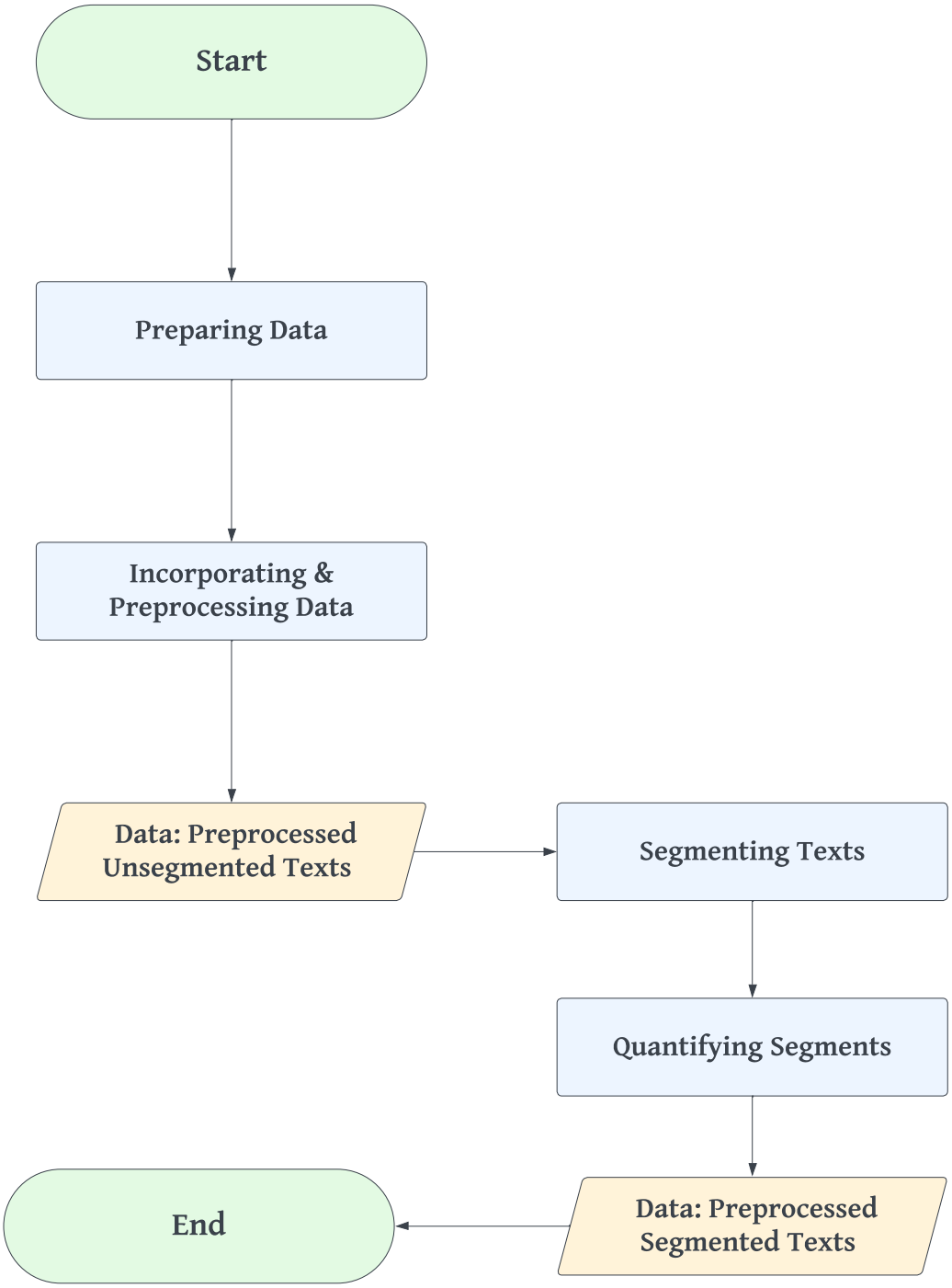


Figure 2.1: Data Preparation And Preprocessing Workflow.

In the first phase of this process, which is related to **O1**, the objective was to obtain the data by manually downloading the subject literature defined in Table 1.1 from the *Project Gutenberg* website [2] into text files and then cleaning the text files from unrelated data in the books, such as information related to the *Project Gutenberg*, and not to the contents of the books. In the second phase, the objective was to prepare the data in a tabular format by reading the files and processing them as data frames. Consequently, the latter step

[2] `https://www.gutenberg.org`

involved incorporating additional information into the data frames, such as the genres, the publication date information, and the rating from the *Goodreads* website [3].

The subsequent step was to clean the text to a minimum level by normalizing it, i.e., formatting it to lowercase, converting numbers into textual representation, and removing redundant white spaces. This phase aimed to create a preprocessed dataset covering the entire set of subject books, which was used in the next stage to generate a second dataset specifically focused on the text segments of the books. The following phase is related to **O2**. The study that used stylometry analysis for document classification partitioned the text for its task into text components, each consisting of ten overlapped sentences [11]. However, this study segmented the books into 15 consecutive sentence segments with five overlapping sentences from the preceding segment; overlapping facilitates maintaining context in the text segments with the preceding and succeeding segments. As for the size of the segment, given that *RoBERTa* can process sequences of a maximum length of 512 tokens a sequence [6][15][11]. Having the text segments maintain the size of 15 sentences is suitable at this stage; making the segment size more minor may lead some segments to lose contextual information. Similarly, making the segment size greater than 15 may lead some segments to be larger than 512 tokens. Using overlapping segmentation would maintain context by ensuring that the boundaries of the segments are not arbitrarily split, preserving the contextual information in the segments [23][24][15].

After segmenting the books into segments, the last stage of this process involved computing the text segments' textual properties. For the textual properties, the text segments underwent stylometry analysis using criteria similar to the initial study, adding to it *Sentiment Polarity*. Table 2.1 provides a detailed list of the stylometry metrics used in the study [11]. While various stylometry metrics exist, like Distance-Based measures such as *Burrow's Delta*, *Kullback-Leiber Divergence Method*, or statistical measures like *Zipf's Law*, *Herdan's C*, *Guiraud's R* [24], this study only applied the metrics that were used in the work of *H. Elahi and H. Muneer* [11] in the authorship attribution task to evaluate this approach for this task using the same metrics. The study used *Sentiment Polarity* in combination with the other metrics to analyze the literature linguistics properties, meaning *Sentiment Polarity* was not in the scope of the authorship attribution task. Accordingly, this phase concluded after computing the stylometric computations, producing the second dataset containing the text segments and their associated textual characteristics.

[3] `https://www.goodreads.com`

Table 2.1: Stylometry Metrics [24][11][25][26] [27][28][29][30].

Metric	**Formula**
Average Functional Words	$\frac{\text{Total Number of Functional Words in the Text}}{\text{Total Words}}$
Average Punctuation	$\frac{\text{Punctuation}}{\text{Total Words}}$
Average Sentence Length By Char	$\frac{\text{Total Characters}}{\text{Total Sentences}}$
Average Sentence Length By Word	$\frac{\text{Total Words}}{\text{Total Sentences}}$
Average Special Characters	$\frac{\text{Total Number of Special Characters in the Text}}{\text{Total Words}}$
Average Syllables Per Word	$\frac{\text{Total Number of Syllables in the Text}}{\text{Total Words}}$
Average Word Length	$\frac{\text{Total Characters of Words}}{\text{Total Words}}$
Brunet's Measure W	$N^{Voc^{-a}}$
Dale-Chall Readability	$0.1579\left(\frac{\text{Difficult Words}}{\text{Total Words}} \times 100\right) + 0.0496\left(\frac{\text{Total Words}}{\text{Total Sentences}}\right)$
Flesch Reading Ease	$206.835 - 1.015\left(\frac{\text{Total Words}}{\text{Total Sentences}}\right) - 84.6\left(\frac{\text{Total Syllables}}{\text{Total Words}}\right)$
Flesch-Kincaid Grade Level	$0.39\left(\frac{\text{Total Words}}{\text{Total Sentences}}\right) + 11.8\left(\frac{\text{Total Syllables}}{\text{Total Words}}\right) - 15.59$
Gunning Fog	$0.4\left[\left(\frac{\text{Total Words}}{\text{Total Sentences}}\right) + 100\left(\frac{\text{Complex Words}}{\text{Total Words}}\right)\right]$
Hapax DisLegemena	$\frac{Voc_2(T)}{N}$
Hapax Legomena	$\frac{\lvert Voc_1(T)\rvert}{\lvert Voc(T)\rvert}$
Honore Measure R	$100 \times \log\left(\frac{N}{1-\frac{V_1}{V}}\right)$
Sentiment Polarity[4]	$\frac{\sum_{i=1}^{n} \text{Polarity}(w_i)}{n}$
Shannon Entropy	$-\sum_{i=1}^{n} p_i \log_2(p_i)$
Sichel's Measure S	$\frac{\lvert Voc_2(T)\rvert}{\lvert Voc(T)\rvert}$
Simpson's Index	$\sum_{r=1}^{r} \frac{r}{n} \cdot \frac{r-1}{n-1} \cdot \lvert Voc_r(T)\rvert$
Type-Token Ratio	$\frac{\lvert Voc(T)\rvert}{n}$
Yule's Characteristic K	$10^4 \cdot \left[-\frac{1}{n} + \sum_{r=1}^{r} \frac{r}{n} \cdot \frac{r}{n} \cdot \lvert Voc_r(T)\rvert\right]$

In terms of the defined stylometry metrics in Table 2.1, these metrics provide various textual analyses regarding the characteristics of a text, which, in turn, would yield information about the style of the writer of the text in question. In essence, writing styles are subjective, meaning authors have distinctive literary and language manners to communicate ideas, views, thoughts, stories, and imaginations in written form. In other words, a writing style encompasses different characteristics, including but not limited to structure, vocabulary, punctuation, and language use [24][31][32]. For *Structural and Lexical* features, the following metrics apply: *Average Functional Words* provides insight into the linguistic complexity of a text, *Average Sentence Length By Char* and *Average Sentence Length By Word* offer insight into the structure complexity. As for *Average Special Characters*, it measures the average use of special characters, including $, in a text, *Average Punctuation Count*, measures the average punctuation in a text. Regarding *Average Syllables Per Word*, it helps assess readability, and *Average Word Length* is beneficial in evaluating the sophistication of vocabulary in a text [24][11].

When it comes to *Vocabulary Richness*, *Hapax Legomenon* focuses on unique words in a given text, providing information about vocabulary richness. However, unlike *Hapax Legomenon*, *Hapax DisLegemena* focuses on the words that occur twice in a text. *Honores R Measure* also measures vocabulary richness by considering the vocabulary, the unique words, and the total number of words. *Sichel's Measure S* measures the number of words that occur twice in a text to the total number of words that occur once in the text. Additionally, *Brunets Measure W* measures the vocabulary diversity of a text. When it comes to *Yules Characteristic K*, it is a statistical metric that measures the degree of linguistic diversity in a text. However, *Shannon Entropy* measures the randomness in s text considering the vocabulary and complexity of the text structure. Moreover, *Simpson's Index* measures the diversity in a text. *Type-Token Ratio* measures the lexical diversity [24][11][33].

As for *Readibility*, *Flesch Reading Ease*, measures the readability score in a text, where the scores span from 1 to 100. Similarly, *Flesch-Kincaid Grade Level* measures the readability score on a scale corresponding to the U.S. educational system. *Gunning Fog Index* measures a score that reflects the number of years of formal education a reader requires to understand the text on first reading [25][26][33].

Regarding *Sentiment Polarity*[5], it measures the text's tone to determine whether its sentiment is positive, negative, or neutral [24].

2.2 Data Analysis

In this process phase, which is related to **O3**, it analyzed the preprocessed dataset for the text segments to asses its balance, meaning it analyzed the distribution of the text segments per author by computing the mean and the standard deviation to determine its balance [34].

2.3 Literature Analysis

In this process phase, which is related to **O4**, the objective was to extract the numerical features from the text segments dataset for the subject books and normalize them using the *Z-Score Normalization* to have zero mean and unit variance to mitigate the influence of differing scales during clustering - the formula for *Z-Score Normalization* is as follows,

[5] `https://textblob.readthedocs.io/en/latest/`

where X is the original data values, μ is the mean of the data, and σ is the standard deviation of the data [35][34]:

$$Z = \frac{X - \mu}{\sigma}$$

Following standardization, the subsequent step was performing an exploratory analysis by executing the *K-Means* clustering algorithm and using inertia as a metric of the cluster quality on different values of k for a range of values spanning from one to nine to obtain empirical data to determine the optimal number of clusters. Accordingly, after obtaining the optimal k for the number of clusters, the *K-Means* clustering algorithm was applied with the obtained k value. Then, the features within each cluster were computed by calculating their mean values to understand the characteristics of the clusters. For visualization, *t-SNE* was employed to reduce the high-dimensional clustered data points in two-dimensional space.

Lastly, this phase concluded by evaluating the linear relationships between the pairs of the numeric features, i.e., the textual properties of the text segments using *Pearson Correlation Coefficient* and then identifying the pairs of features with a high correlation coefficient that is more significant than the absolute value of 0.75. The formula for the *Pearson Correlation Coefficient* of two variables is as follows, where it divides the covariance between the two variables X and Y, by the standard deviations of X and Y [36]:

$$\rho_{(X,Y)} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

2.4 Authorship Classification

This section uses different models and classification magnitudes to present the authorship classification task associated with **O5** and **O6**. Figure 2.2 illustrates the process of this phase from a general point of view.

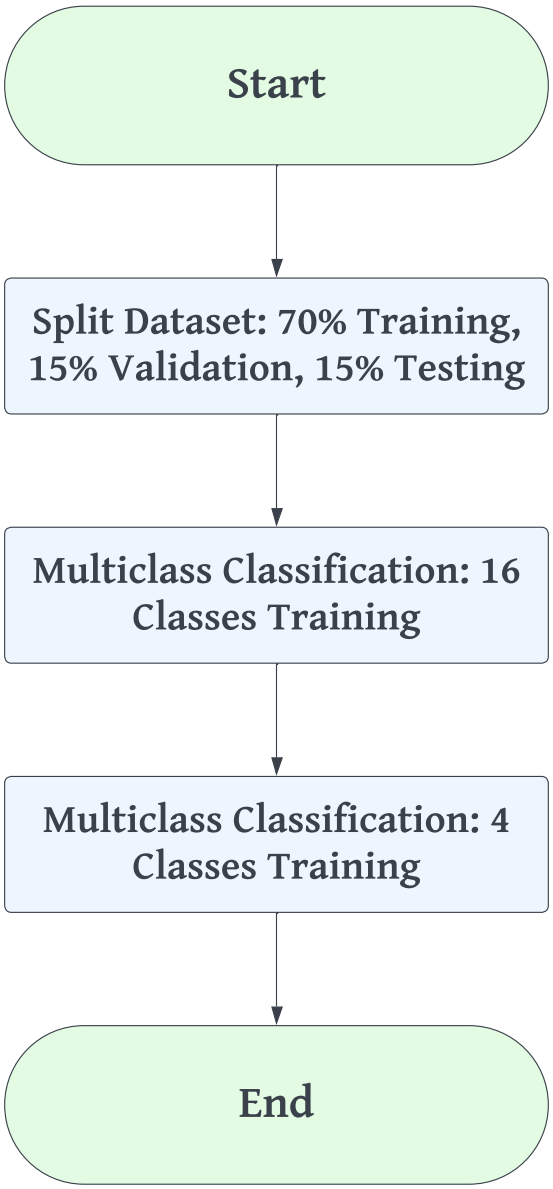


Figure 2.2: The Multiclass Classification Process.

For *LightGBM* and *TabNet*, the process used the segmented texts' features, i.e., the computed textual characteristics, as the descriptive features for training. It normalized the textual characteristics using the zero mean and unit variance [35]. However, for *RoBERTa*, the process used the text segments instead as the descriptive features for training. As for the target features, the process encoded the authors' names as the target features and used them for all models. Principally, the process phase ran one primary training setting using one dataset, dividing and stratifying the data into splits of 70% for the training set, 15% for the validation set, and 15% for the test set.

There are three books per author, and the size of the books is diverse, which means that some books produced considerably more text segments than others, which could have potentially imbalanced the dataset. Therefore, for each subject model, two training subprocesses were conducted for multiclass classifications with varying sizes of classes,

16 and 4. These different training subprocesses were undertaken to obtain a combination of results for any given model to compare the results, evaluate the models' performance, and assess the methodology reasonably in the event of imbalanced dataset, overfitting, small dataset, or high variance in the dataset.

The first subprocess training was classifying an author out of the 16 authors, meaning training an instance of any given model of the defined models for all classes. In the second subprocess, the adjustment was downsizing the data to only the top four authors with the most text segments to train the same model on this modified set to classify an author out of these four, implying training the model in question only on four classes. Consequently, two instances were trained separately during the primary process for any given model.

2.4.1 LightGBM Classification

In this task, the study utilized the optimization framework *Optuna*[6] by running 50 trials to find the model's optimal parameters from a defined parameters settings listed in Table 2.2 focusing on *Accuracy* and *Validation set* as the metrics for evaluating performance.

Table 2.2: Parameter Definition for LightGBM Tuning

Parameter	**Values**
N Estimators	50 to 200
Max Depth	3 to 10
Min Child Samples	20 to 100
Learning Rate	0.005 to 0.05
Subsample	0.6 to 0.9
Number of Leaves	20 to 50

2.4.2 TabNet Classification

For the learning settings of *TabNet*, the model setup entailed setting the optimizer to *Adam*, with the *Learning Rate* set to 0.02, and configuring the *Learning Rate* scheduler that adjusts the *Learning Rate* every ten *epochs* by multiplying it with a factor of 0.9 to adjust the *Learning Rate* during training. Regarding the *Sparse Attention Mechanism*, the model used *entmax* to aim for better performance. In terms of the model training, the settings for the training encompassed using *Accuracy* and *Validation set* as the metrics for evaluating performance, scoping the training to a maximum of 100 *epochs* with the patience at three to allow the training to stop at an early stage if there is no improvement after three epochs or to prevent overfitting.

2.4.3 RoBERTa Classification

The learnings settings for *RoBERTa* entailed setting the *Warmup Steps* to 50, using *Polynomial* as a decay scheduler, and implementing an early stopping mechanism with patience three to prevent the model from overfitting or to stop the learning process if the learning does not improve. In terms of the model training, the settings for the training encompassed using *Accuracy* and *Validation set* as the metrics for evaluating performance.

[6] `https://optuna.org`

2.5 Validation

The study considered validation in the early stages and integrated it into each process phase as a systematic workflow of the entire study. The latter used various technologies, methods, and literature reviews to minimize errors.

As for the data preparation and preprocessing, the study executed this phase systematically to minimize inaccuracies. This was achieved using advanced *Python* libraries, each serving a specific purpose, such as cleaning the text, preprocessing, and computing the textual features defined in Table 2.1. The libraries used included but were not limited to: *TextBlob*, *Natural Language Toolkit (NLTK)*, *Textatistic*, *SpaCy*, *Readability*, *Pandas*, and *Numpy*. Hence, this phase depended on the built-in functions and existing tools to serve its objectives for the most part. Despite all efforts, given that the technologies are imperfect, there are sometimes error margins.

Furthermore, after training, the performance of the models was evaluated on the test set, computing the following measurement metrics: *Accuracy*, *Precision*, *Recall*, and *F1-score* to assess the model's performance and look for a balance between these metrics to ensure the model is both accurate and reliable.

When it comes to the *LightGBM* classification task, this task utilized *Optuna* to conduct a comprehensive search over the defined parameters to identify the optimal parameters. As for *TabNet* and *RoBERTa*, both implemented early stopping mechanisms to prevent overfitting and underfitting.

As for literature analysis, the unsupervised approach concerning the clustering using *K-Means* algorithm, the study, as mentioned earlier, has preprocessed the data using built-in functions to provide accurate data as much as possible and prevent introducing biases and clustering books and authors inaccurately or misrepresent them.

3 Theoretical Background

This section presents the theoretical aspects of the supervised machine algorithms chosen for this study to perform authorship classification tasks from a high-level perspective.

3.1 LightGBM

Light Gradient Boosting Machine is a superviced machine learning algorithm, typically known as *LightGBM*. As the name may suggest, it adopts the mechanism of *Gradient Boosting* algorithm, an optimizer algorithm whose approach is sequential to build a model composed of a combination of models. In mathematical terms, *Gradient Boosting* algorithm is as follows [37]:

$$F_M(x) = F_0(x) + \sum_{m=1}^{M} \gamma_m h_m(x)$$

Gradient Boosting starts its training process with a base model; $F_0(x)$ and improves its learning process dynamically and iteratively until the process concludes according to the initial configurations, where each iteration; m, the algorithm adds a new model, $h_m(x)$ with updated learning settings; γ_m based on the error margin of the preceded model to the sequence of the previous models to minimize the error margin and to optimize the performance by reducing the overall margin of error between the actual values and the expected values [37][38].

Because of *Gradient Boosting*'s sequential process, it can go through the features of the training set to construct a model that can capture the importance of the features in the training set and achieve high performance. Logically, given its architectural nature, which is additive, where it continuously adds a new model to its sequence of models to produce the final model, which can handle the task at hand gracefully, the process can be demanding in resources and complex. In terms of resources, *Gradient Boosting* algorithm can be demanding when the data in question is significant or in high dimensionality, be it time or computational resources, as it can consume excessive computational resources depending on the dimensionality and the size of the training set, which makes it ineffective in some situations[37][38].

Consequently, the limitations of the *Gradient Boosting* in terms of scalability and efficiency have led to the inception of *LightGBM* to address these limitations. Regarding *LightGBM*, it is a tree-based algorithm that applies the *Gradient Boosting* mechanism in a fashion that applies two strategies. The first strategy is a *Gradient-Based One-Side Sampling (GOSS)* strategy, where the algorithm focuses on the most significant error margin to split the data into two leaves, meaning the algorithm will split the groups into two leaves that are somewhat similar in terms of the error margin. Iteratively, it selects the next leaf with the highest error margin, groups the data points and divides them into two leaves, updates the model learning settings, and repeats the same process until it concludes. The second strategy *Exclusive Feature Bundling (EFB)*, which occurs before the learning process, is to reduce the dimensionality of the training data by analyzing the features of the data to encapsulate the features that do not change exclusively simultaneously [37][38][39].

With the approach of the *LightGBM* algorithm, the focus is on the data points that yield the most information gain, meaning it does not need to scan all features to discover the importance of the features of the given dataset or focus on the data points with insignificant error margin. Furthermore, it is scalable as it can reduce dimensionality in

data and drop random instances with low error margins, the data points that do not provide any substantial information gain that benefits the learning process. Accordingly, *LightGBM*'s mechanism makes it efficient, flexible, high performer, and scalable to handle large datasets. Regardless, *LightGBM* still imposes some challenges. Considering its complex structure, which is tree-based, interpretability may not be an intuitive task to understand how it reasons to make its predictions. Furthermore, it may require optimal tuning to achieve high performance, which may require a deep understanding of how the parameters influence the learning process, and it may also introduce the risks of overfitting when the dataset is relatively small [37][38][39].

3.2 TabNet

In 2019, *Google Cloud AI* researchers introduced the algorithm *TabNet*, which is a *Deep Learning* algorithm. One of the benefits of *TabNet* is that it does not require extensive data preprocessing; on the contrary, it requires minimal data preprocessing. Regarding the data format, *TabNet* primary objective is to process data in tabular format [40].

The approach of *TabNet* is similar to *Decision Trees*; it builds its decision sequentially to reach the final solution. However, *TabNe* processes the data iteratively where at each it applies the *Attention Mechanism*, which analyzes the data's features in the current iteration using the obtained information from the previous step to determine the features' importance at the current state. After examining the importance of the features, *TabNet* performs the *Sparse Feature Selection Mechanism* by prioritizing and selecting a limited subset of the features with high importance to process them through the model's neural network layers and reach a decision about the feature impact on the outcome. Consequently, based on the gained decision, it updates its learning settings and *Attention Mechanism* to proceed to the next step and repeats the same process until it concludes. Once the process finishes its iterations, it uses all decisions to construct the final model [40].

TabNet provides multiple benefits, one of which is handling flexibly high dimensional datasets through its *Attention Mechanism* and *Sparse Feature Selection Mechanism*, meaning it processes features selectively rather than scanning all simultaneously. Additionally, given its selective approach, it does not make the model complex, which can provide an understanding of how the algorithm reasons in its decisions. Another advantage is that *TabNet* is a type of *Neural Network* model, which allows it to handle datasets that maintain non-linear relationships gracefully. However, *TabNet* may be demanding in terms of computation resources given its *Neural Network* architecture and may not work well on imbalanced data or small data [41][40].

3.3 RoBERTa

As for *Natural Language Processing*, the model *Robustly Optimized BERT Approach*, typically known as *RoBERTa*, is a pretrained *State-Of-The-Art* model that processes natural human language in text format and it is a type of *Bidirectional Encoder Representations from Transformers*, known as *BERT* [15].

Google developed the language model *BERT*, which adopts the *Transformer* architecture, a type of *Neural Network* architecture that can analyze complex data structures [42]. As for the *Transformer* architecture, which relies on the *Attention Mechanism* and is composed of two components: *Encoder* and *Decoder*. The task of *Encoder* is to use the *Attention Mechanism* to process the input, which is a sequence of words, to embed the words through a series of layers, to present them in a vector form, maintaining the contextual information of the provided sequence, each vector injected with additional

information referencing the positions of the words in the sequence. For the *Encoder*, its task is to process the output of the *Encoder* using the *Attention Mechanism* to generate the output sequentially, meaning it sequentially constructs its final output by predicting the next word in the output sequence considering the encoded information and the previously generated words [6][43].

BERT undergoes two stages: the pretraining stage for learning and the fine-tuning stage for configuring it to perform a specific task. In simple terms, *BERT* applies the *Transformer* architecture, which means it contains an encoder, the *Encoder* of *BERT*, as the name may suggest, it operates in both directions simultaneously where the algorithm masks a partition of the sequence input, processes the information from left-right and right-left to understand the contextual information of the sequence efficiently to predict the masked portion. Regarding the sequence length, *BERT* can process a 512-token sequence at a time [6][43].

For *RoBERTa*, it is an optimized version of *BERT*, which was pretrained on a larger corpus than *BERT*, and unlike *BERT*, it does not involve prediction in the training like *BERT*, instead it focuses on discovering and understanding the relationships in the corpus. Similarly, it can process a 512-token sequence at a time [6][15][43].

4 Results and Analysis

This section presents and analyzes the results.

4.1 Data

The total number of tokens in all books is approximately 8,922,516, the total number of text segments yielded from segmenting the subject books is 28,836, and the average number of tokens in a text segment is approximately 309.4. Figure 4.1 shows the text segmentation across the 16 authors.

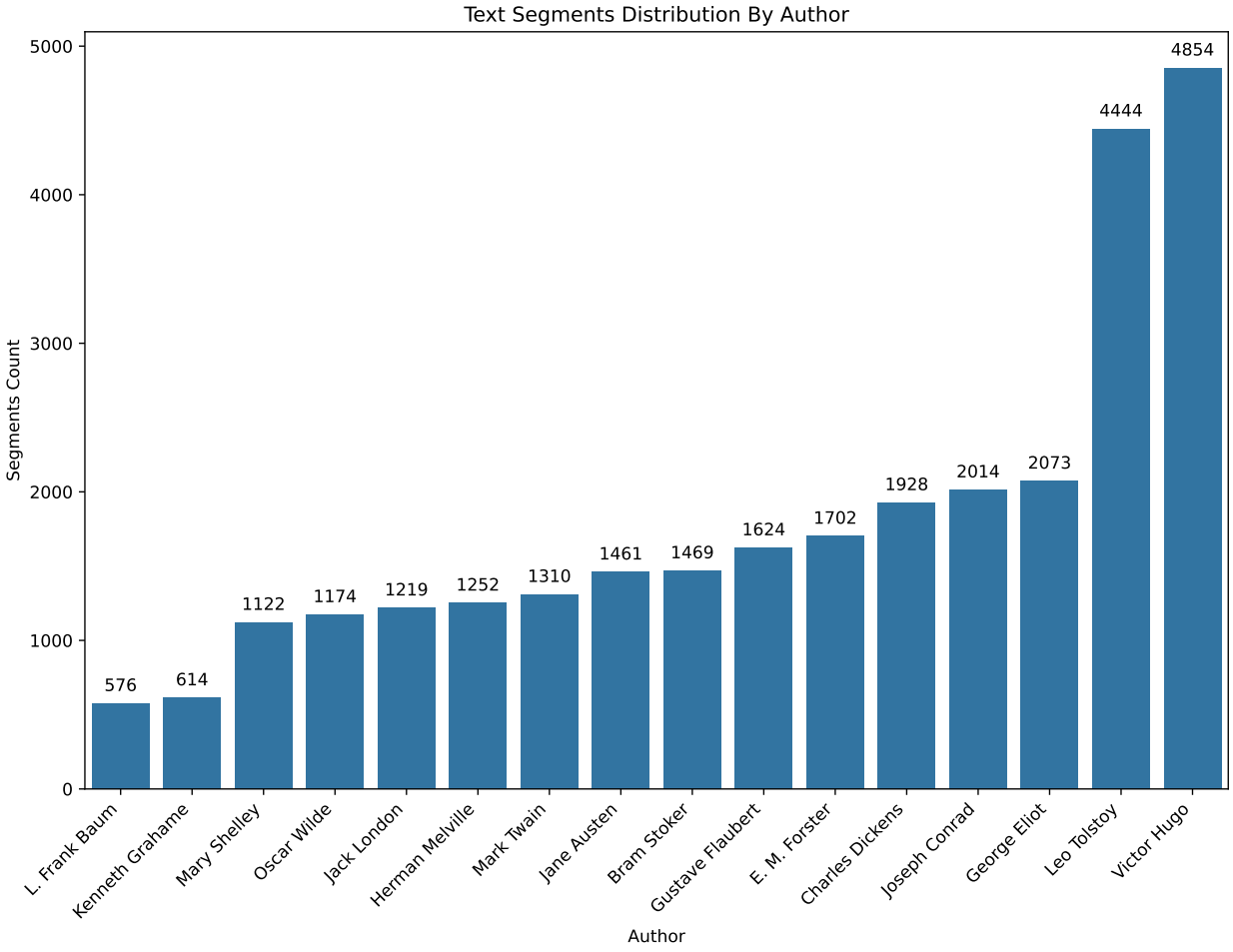


Figure 4.1: Distribution Of Text Segments Across All Authors.

Regarding the proportionality of the text segments, Figure 4.2 shows the portion of text segments per author. The most significant portion goes to the author *Victor Hugo*, 16.8%, and the most minor portion belongs to the author *L. Frank Baum*, which is 2%.

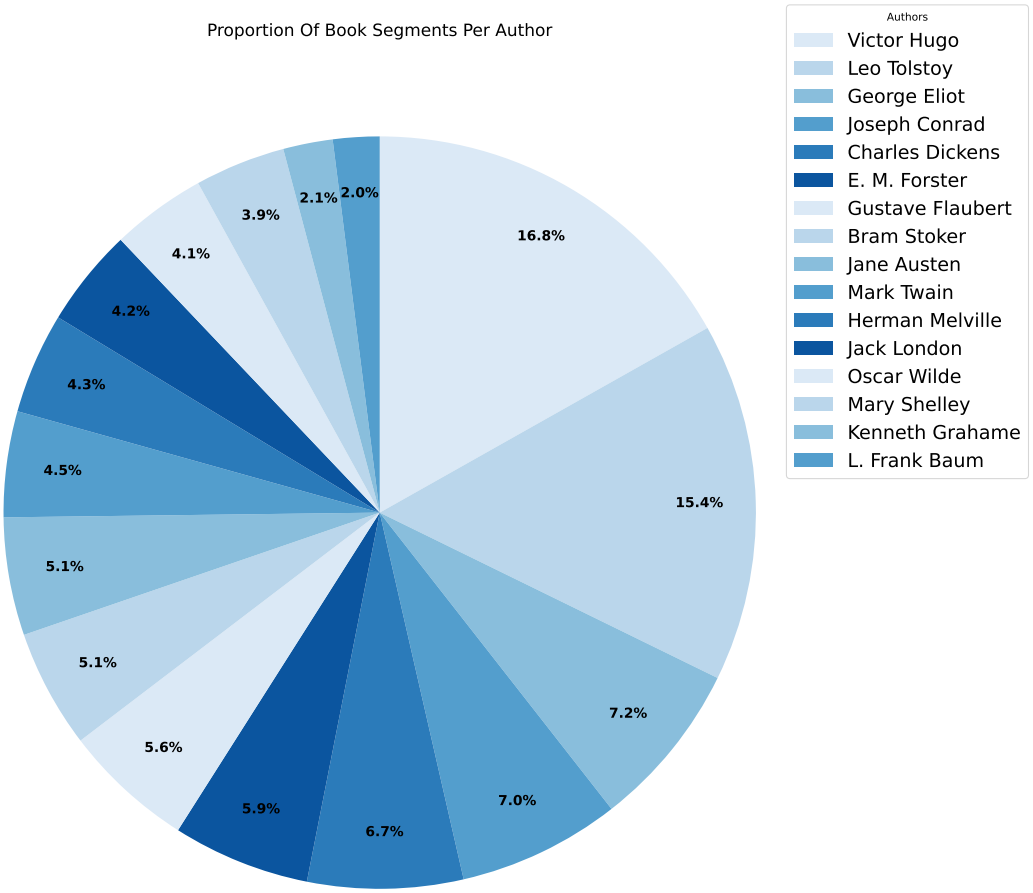


Figure 4.2: Proportion of Book Segments Per Author.

The mean of the text segments per author is as follows [34]:

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

N is the total number of authors and x_i is the number of text segments per author. Substituting the values:

$$\bar{x} = \frac{\begin{array}{c}(576 + 614 + 1122 + 1174 + 1219 + 1252 + 1310 + 1461 + 1469 + \\ 1624 + 1702 + 1928 + 2014 + 2073 + 4444 + 4854)\end{array}}{16}$$

$$\bar{x} = \frac{28836}{16} = 1802.25 \approx 1802$$

The computation above shows that 1802, is approximately the mean number of text segments per author. With 16 authors, the yielded mean value is the ideal number of text segments per author to maintain a balanced distribution in the dataset. Figure 4.3 illustrates the deviation from the mean value across the 16 authors.

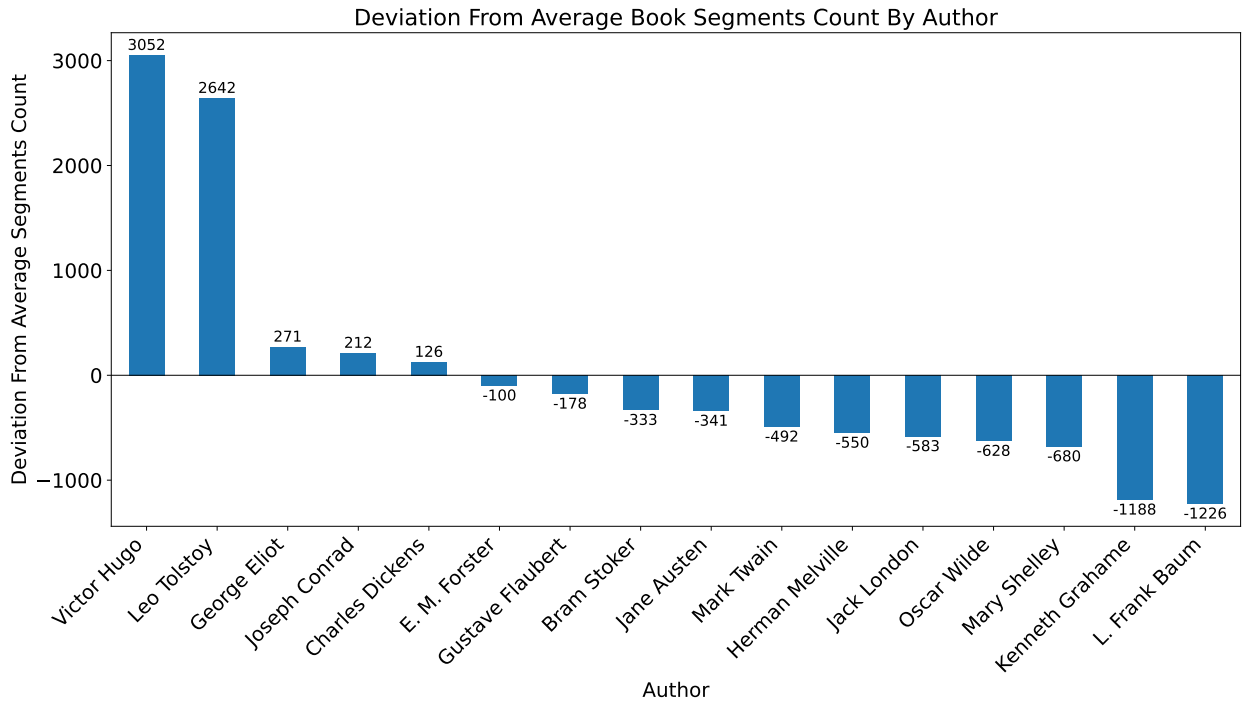


Figure 4.3: Deviation From The Average Segment Counts By Author.

For the standard deviation, the computation is as follows [34]:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$$

Substituting the values:

$$\sigma = \sqrt{\frac{(576 - 1825.25)^2 + (614 - 1825.25)^2 + \ldots + (4854 - 1825.25)^2}{16}}$$

$$\sigma \approx 1155.25$$

The results indicate substantial differences and variability in the number of text segments across the authors in the dataset. In addition, the dataset may be small for the multiclass classification task involving 16 authors; enhancing the number of books per author may be required to make the models for the multiclass classification tasks able to capture the complexity of the textual properties properly to distinguish an author from the 16 authors.

According to the results, the proportion of the used text segments of the top four authors; *Victor Hugo*, *Leo Tolstoy*, *George Eliot*, and *Joseph Conrad* for multiclass classification tasks of 4 authors is slightly less than half of the entire dataset; the calculations are as follows:

$$16.8\% + 15.4\% + 7.2\% + 7.0\% = 46.4\%$$

$$4854 + 4444 + 2073 + 2014 = 13,385$$

As for the division of the preprocessed dataset, Table 4.1 shows the dataset splitting according to the multiclass classification tasks.

Table 4.1: Data Splits For Multiclass Classification Tasks.

Table 4.2: (16 Authors).

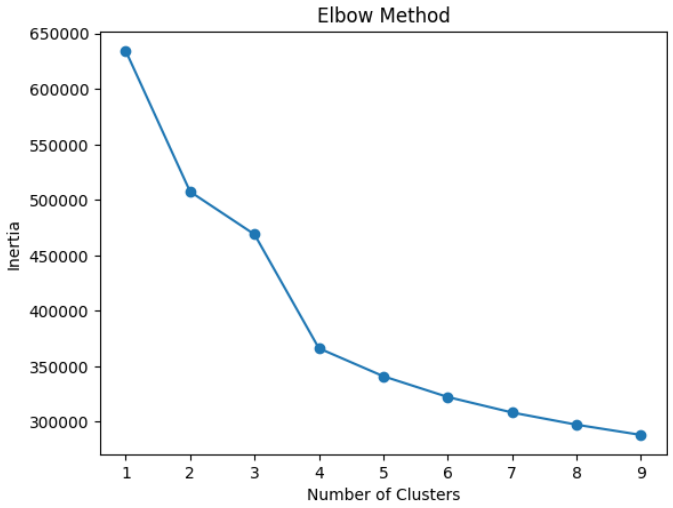
Dataset	**Rows**	**Percentage**
Training set	20185	$\approx 70\%$
Validation set	4325	$\approx 15\%$
Test set	4326	$\approx 15\%$
Total	28,836	100%

Table 4.3: (4 Authors).

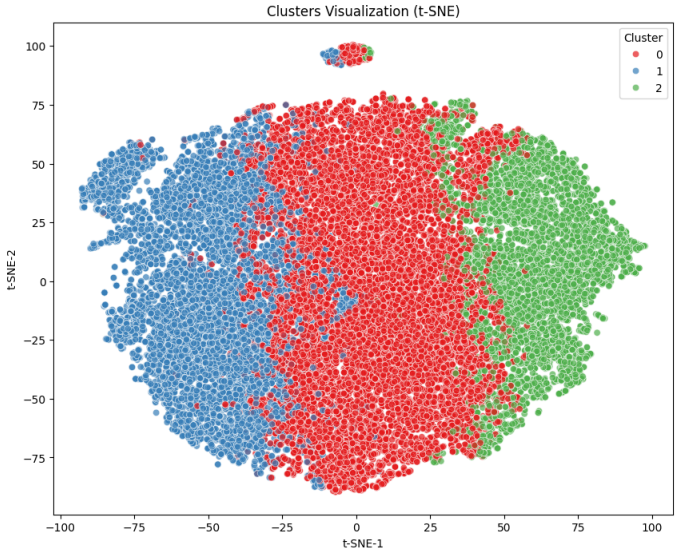
Dataset	**Rows**	**Percentage**
Training set	9369	$\approx 70\%$
Validation set	2008	$\approx 15\%$
Test set	2008	$\approx 15\%$
Total	13,385	100%

4.2 RQ1

This section presents and analyzes the results related to **RQ1**. Figure 4.4a demonstrates the optimal number of clusters for the *K-Means* clustering algorithm, which is 3, and Figure 4.4b visualizes the formation of the clusters.



(a) The Optimal Number Of Clusters $k = 3$.



(b) Clusters Visualization Based On Textual Features.

Figure 4.4: The Clusters Of The Text Segments.

Table 4.4 contains the definitions of the acronyms used in Table 4.5 and 4.6.

Table 4.4: Acronym Definitions.

Acronym	Definition
AFW	Average Functional Words
AP	Average Punctuation
AR	Average Rating
ASC	Average Special Characters
ASLC	Average Sentence Length By Char
ASLW	Average Sentence Length By Word
ASPW	Average Syllable Per Word
AWL	Average Word Length
BM	Brunets Measure W
C	Cluster
DCR	Dale-Chall Readability
FKGL	Flesch-Kincaid Grade Level
FRE	Flesch Reading Ease
GF	Gunning Fog
HD	Hapax DisLegemena
HL	Hapax Legomena
HM	Honore Measure R
S	Sentiment Polarity
SE	Shannon Entropy
SI	Simpson's Index
SM	Sichele's Measure S
TTR	Type-Token Ratio
YCK	Yule's Characteristic K

Table 4.5 and 4.6 provide the characteristics of the book clusters in terms of their textual properties.

Table 4.5: Textual Properties Across Clusters.

C	AR	AWL	ASLW	ASPW	ASC	AP	AFW	TTR	HM	HL	SM
0	3.91	5.67	20.85	1.72	0.003	0.028	0.12	0.52	740.62	0.001	0.094
1	3.94	5.20	12.66	1.58	0.005	0.04	0.11	0.55	682.8	0.003	0.098
2	3.88	6.12	31.78	1.87	0.002	0.02	0.12	0.48	785.54	0.001	0.1

Table 4.6: Textual Properties Across Clusters.

C	HD	YCK	SI	BM	SE	FRE	FKGL	DCR	GF	ASLC	S
0	0.002	754.349	0.925	5.518	4.226	76.461	7.087	23.948	9.325	114.692	0.08
1	0.005	755.302	0.924	5.793	4.272	86.595	4.069	21.701	6.223	68.341	0.071
2	0.002	748.828	0.925	5.467	4.208	60.916	11.755	27.123	14.068	180.035	0.091

For the book clustering, Table 4.7 provides information about the grouped books into the three different clusters.

Table 4.7: The Books Per Clusters.

Cluster	Book	Author	Rating
0	A Passage To India	E. M. Forster	3.68
0	Anna Karenina	Leo Tolstoy	4.09
0	Dracula	Bram Stoker	4.01
0	Heart Of Darkness	Joseph Conrad	3.43
0	Madame Bovary	Gustave Flaubert	3.70
0	Ninety- Three	Victor Hugo	4.08
0	Sentimental Education	Gustave Flaubert	3.81
0	The Adventures Of Huckleberry Finn	Mark Twain	3.83
0	The Call Of The Wild	Jack London	3.90
0	The History Of A Crime	Victor Hugo	3.66
0	The Lair Of The White Worm	Bram Stoker	2.78
0	The Marvelous Land Of Oz	L. Frank Baum	3.80
0	The Wind In The Willows	Kenneth Grahame	4.01
0	The Wonderful Wizard Of Oz	L. Frank Baum	4.00
0	War And Peace	Leo Tolstoy	4.16
0	White Fang	Jack London	4.02
1	A Room With A View	E. M. Forster	3.90
1	Howards End	E. M. Forster	3.96
1	Lady Windermeres Fan	Oscar Wilde	3.92
1	Moby- Dick	Herman Melville	3.54
1	The Adventures Of Tom Sawyer	Mark Twain	3.92
1	The Importance Of Being Earnest	Oscar Wilde	4.17
1	What Men Live By	Leo Tolstoy	4.10
2	A Connecticut Yankee In King Arthurs Court	Mark Twain	3.77
2	A Tale Of Two Cities	Charles Dickens	3.87
2	Bartleby	Herman Melville	3.93
2	Dream Days	Kenneth Grahame	3.81
2	Emma	Jane Austen	4.05
2	Frankenstein	Mary Shelley	3.87
2	Great Expectations	Charles Dickens	3.79
2	Les Miserables	Victor Hugo	4.20
2	Lord Jim	Joseph Conrad	3.62
2	Middlemarch	George Eliot	4.02
2	Nostromo	Joseph Conrad	3.81
2	Oliver Twist	Charles Dickens	3.88
2	Ozma Of Oz	L. Frank Baum	3.96
2	Pride And Prejudice	Jane Austen	4.29
2	Salammbo	Gustave Flaubert	3.74
2	Sense And Sensibility	Jane Austen	4.08
2	Silas Marner	George Eliot	3.68
2	The Golden Age	Kenneth Grahame	3.80
2	The Jewel Of Seven Stars	Bram Stoker	3.42
2	The Last Man	Mary Shelley	3.37
2	The Mill On The Floss	George Eliot	3.82
2	The Piazza Tales	Herman Melville	3.84

Continued on next page

Table 4.7 – *Continued from previous page*

Cluster	Book	Author	Rating
2	The Picture Of Dorian Gray	Oscar Wilde	4.12
2	The Sea Wolf	Jack London	4.05
2	Valperga	Mary Shelley	3.54

Table 4.5 and 4.6 encompass the results obtained using clustering and the techniques defined in Section 2.3, where each cluster serves a distinct group of books with style, complexity, and readability variations.

For **Cluster 0**, the number of words per sentence is approximately 20.85, and the number of syllables per word is about 1.72, pointing to a certain complexity. The Type-Token Ratio is approximately 0.52, and *Honore Measure R* is around 740.62, denoting a balanced vocabulary richness without high complexity. A *Flesch Reading Ease* score of 76.46 and a *Gunning Fog* Index of 9.33 show that these texts are relatively easy to understand.

Cluster 1 contains the shortest average word length of 5.20 and sentence length of 12.66, which implies that the texts' structures are highly readable and straightforward. It maintains a higher punctuation count of 0.040, which may entail dynamic writing. In addition, it scores the highest *Flesch Reading Ease* score, 86.59, and the lowest *Dale-Chall Readability* score, 21.70, conveying that these texts might be structured to be easily understood by a wide range of audiences.

Regarding **Cluster 2**, the results establish that this cluster holds the longest words, which are 6.12 characters per word on average, and the longest sentences, 31.78 words per sentence, hinting at the use of specialized vocabulary and complex sentence structures. Despite a low *Type-Token Ratio* of 0.4, high scores in metrics like *Honore Measure R* 785.5 and low *Hapax Legomenon* 0.001 indicate a focus on a specific vocabulary used repeatedly. In addition, it scored the lowest *Flesch Reading Ease* score, 60.92, and the highest *Flesch-Kincaid Grade Level*, 11.75, demonstrating that these texts might be for readers with a high level of education. It is worth mentioning that the sentiment scores among the clusters are all close to one other, ranging from 0.07 to 0.09, indicating that the clusters have relatively neutral overall sentiment.

Additionally, the books, on average, contain similar characteristics; for instance, the *Clusters* maintain more or less similar scores for *Average Functional Words*, *Shannon Entropy*, and *Simpson's Index*, indicating that computing all of these metrics may not be optimal instead focusing on a smaller set of metrics that yield more information gain and is less computationally demanding may be a reasonable consideration for improvements.

The key findings indicate that **Cluster 0** balances complexity and lexical diversity, meaning moderately complex, while **Cluster 1** maximizes readability and engagement. Although **Cluster 1** maintains small sentence structures, this cluster holds a relatively high *Type-Token Ratio* and *Honore Measure R*, representing a good balance between simplicity and vocabulary richness. Furthermore, **Cluster 1** achieved the highest *Average Rating*, followed by **Cluster 0** and then **Cluster 2**. Lastly, the results illustrate that **Cluster 2** maintains high complexity in structure and vocabulary.

Table 4.8: Top Correlated Features (*Pearson Correlation*).

Feature 1	**Feature 2**	**Coefficient**
Simpson's Index	Yule's Characteristic K	0.999855
Average Sentence Length By Word	Average Sentence Length By Char	0.993611
Average Sentence Length By Char	Dale-Chall Readability	0.993581
Average Sentence Length By Word	Dale-Chall Readability	0.986883
Flesch-Kincaid Grade Level	Gunning Fog	0.983347
Flesch-Kincaid Grade Level	Flesch Reading Ease	0.940416
Flesch Reading Ease	Gunning Fog	0.899913
Honore Measure R	Dale-Chall Readability	0.899297
Honore Measure R	Average Sentence Length By Char	0.893956
Average Sentence Length By Word	Honore Measure R	0.891322

The results in Table 4.8 expose several strong correlation relationships and takeaways. For instance, the *Simpson's Index* and *Yule's Characteristic K* are metrics that measure linguistic diversity within a text. Hence, this correlation is intuitive, given that they both reflect the vocabulary diversity in the texts. For *Average Sentence Length By Word* and *Average Sentence Length By Char*, longer sentences, by word count, are more extended in size by characters. Likewise, this relationship is more or less trivial and straightforward.

Furthermore, the relationship between *Average Sentence Length By Char* and *Dale-Chall Readability* implies that as sentences become longer in size, they typically become more complicated to read, impacting readability scores. The same applies to the *Average Sentence Length By Word* and *Dale-Chall Readability*, noting that longer sentences increase the *Dale-Chall Readability* score, indicating more complex and harder-to-read texts.

Regarding *Flesch-Kincaid Grade Level* and *Gunning Fog*, the metrics assess text readability based on sentence length and word complexity. The *Flesch-Kincaid Grade Level* and *Flesch Reading Ease* indicate that they are inversely related, where a higher *Flesch-Kincaid Grade Level*, meaning complexity, corresponds to a lower *Flesch Reading Ease score*, indicating difficulty. In other words, this correlation points out that as texts become more challenging per the *Flesch-Kincaid* metric, they become relatively less easy to read, according to the *Flesch Reading Ease*.

As for the *Flesch Reading Ease* and *Gunning Fog*, more easy-to-read texts have higher *Flesch Reading Ease* and lower *Gunning Fog* Indices, conveying that they are less complex and more easy-to-read. When it comes to the *Honore Measure R*, vocabulary richness is assessed in the texts. The strong correlation with *Dale-Chall Readability* indicates that texts with richer vocabulary tend to be harder to read, impacting the readability score. *Honore Measure R* and *Average Sentence Length By Char* suggest that the texts with richer vocabularies also tend to have longer sentences, which could contribute to the text's complexity. Similarly, *Average Sentence Length By Word* and *Honore Measure R* indicate that texts with longer sentences use a richer or more complex vocabulary.

The results of the stylometry analysis indicate that textual features can be used to distinguish writing styles. Furthermore, they also show that not all features maintain high correlations, which may suggest that further feature engineering might be needed to exclude the features that may not contribute, especially for other tasks such as classification; instead, they might introduce noise.

Principally, the correlated relationships reflect the following on the clusters of the books: the books in **Cluster 0** show high values in both *Simpson's Index* and *Yule's*

Characteristic K, indicating a rich but complex vocabulary. In contrast, as more easy-to-read texts, the books of **Cluster 1** may maintain a lower correlation between complexity measures *Dale-Chall Readability* and sentence length, as it has shorter sentences and higher readability scores. It might demonstrate less vocabulary diversity than the books of **Cluster 0**. Lastly, the books of **Cluster 2**, with the most complex and least readable texts according to the measurements, likely maintain the strongest correlations between the high readability metrics and long sentences. This cluster shows high values in both *Simpson's Index* and *Dale-Chall Readability*, similar to **Cluster 0** but with a wider extent.

The remaining results associated with this **RQ** are in the appendix, Section A.1.

4.3 RQ2

This section presents and analyzes the results related to **RQ2**.

4.3.1 LightGBM

Regarding the optimization process to find the optimal parameters according to the given task, the *Optuna* Optimization Outputs for both tasks are in Table 4.10 and 4.11. The number of trials for the optimization process for each task is 50. Those parameters are the optimal values for maximizing the model's *Accuracy* for the respective task.

Table 4.9: Optimal Parameters Found.

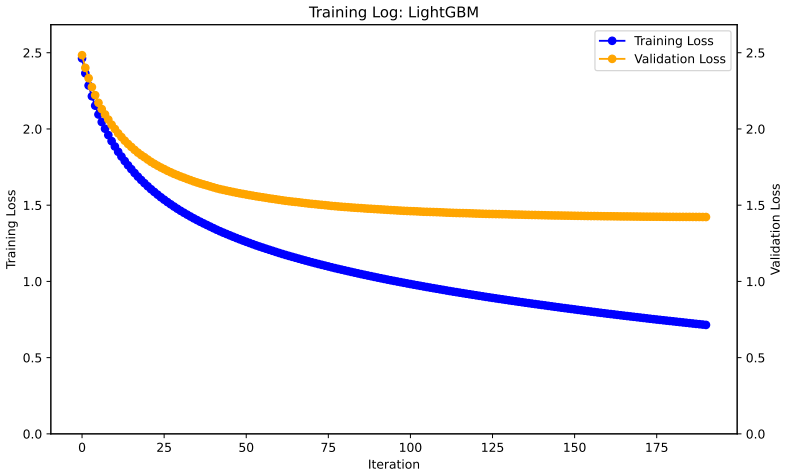
Table 4.10: LightGBM: (16 Authors).

Parameter	Value
N Estimators	191
Max Depth	10
Min Child Samples	29
Number Of Leaves	20
Learning Rate	0.0496
Subsample	0.7891

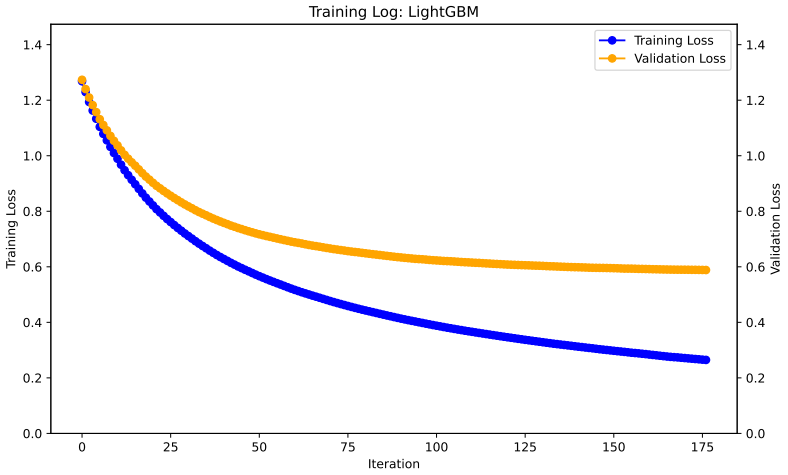
Table 4.11: LightGBM: (4 Authors).

Parameter	Value
N Estimators	177
Max Depth	10
Min Child Samples	39
Number Of Leaves	47
Learning Rate	0.0329
Subsample	0.6774

As for the learning process, Figure 4.5a and 4.5b illustrate the processes for both tasks. In Figure 4.5a, the *Training Loss* starts at approximately 2.5 and gradually decreases and plateaus around 0.7. The *Validation Loss* starts at around 2.5. It decreases initially but then plateaus around 1.5. Regarding Figure 4.5b, the *Training Loss* starts at approximately 1.3 and gradually decreases and plateaus around 0.3. The *Validation Loss* starts at around 1.3. It decreases initially but then plateaus around 0.6, indicating the model generalizes better on the validation set in the multiclass classification task for 4 authors than the 16 authors' task considering the smaller gap between the *Training Loss* and *Validation Loss*.



(a) (16 Authors).



(b) (4 Authors).

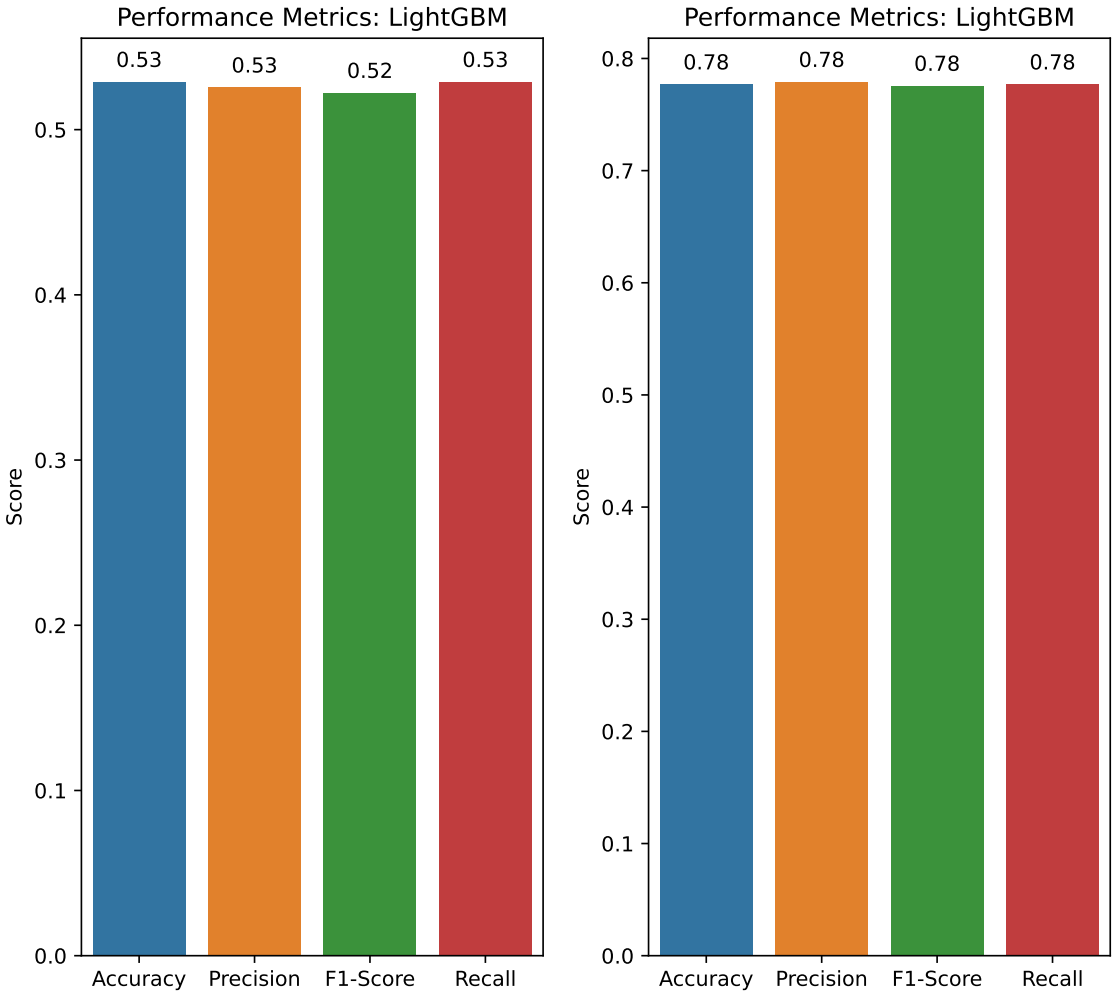
Figure 4.5: Training Progress: *LightGBM*.

For the model's performance in both tasks, Figure 4.6a and 4.6b show the scores according to the metrics. In the multiclass classification task of 16 authors, the model achieves an *Accuracy* of around 53%, implying that it correctly predicts the authors slightly more than half of the time. In the multiclass classification of 4 authors, the model achieves an *Accuracy* of around 78%, signifying that its performance enhanced substantially and became practical. In both tasks, the scores of *Precision*, *F1-score*, and *Recall* are close to one another, offering a balanced performance.

Regarding the performance of the multiclass classification of 16 authors, the performance score is relatively low, implying that further improvements are needed to optimize the performance. The dataset imbalance could have impacted the overall performance. Furthermore, downsizing the magnitude of the multiclass classifications task to 4 authors instead of 16 decreased the dataset's imbalance degree and improved the model's performance. Another aspect is to study downsizing the stylometry metrics to a smaller and more focused set, which would decrease the dimension of the dataset and increase the information gain. Such consideration requires further analysis to filter out the essential metrics.

Considering the character of *LightGBM*, that it requires a large dataset, this also could be a factor that the size of the used dataset is small for 16 authors [37][38][39]. Lastly, another possible improvement is to conduct a post-learning process or comprehensive

hyperparameter-tuning process to find the most optimal parameters from a broader range, which neither was within the scope of this study.



(a) (16 Authors). (b) (4 Authors).

Figure 4.6: Performance Metrics: *LightGBM*.

Figure 4.7 is an Accuracy Matrix belonging to the multiclass classification task of the 16 authors that displays the model *Accuracy* performance per author. The model performed better on *Leo Tolstoy* and *Victor Hugo*, where it achieved the highest *Accuracy* and with *Kenneth Grahame* and *L. Frank Baum*, the model achieved the lowest *Accuracy*. Understandably, *Leo Tolstoy* and *Victor Hugo* have the top number of segments across all authors, and *Kenneth Grahame* and *L. Frank Baum* maintain the lowest number of text segments among all authors. Furthermore, the model misclassified the work of *Leo Tolstoy* for *Victor Hugo* 64 times, and it misclassified the work of *Victor Hugo* for *Leo Tolstoy* 71 times.

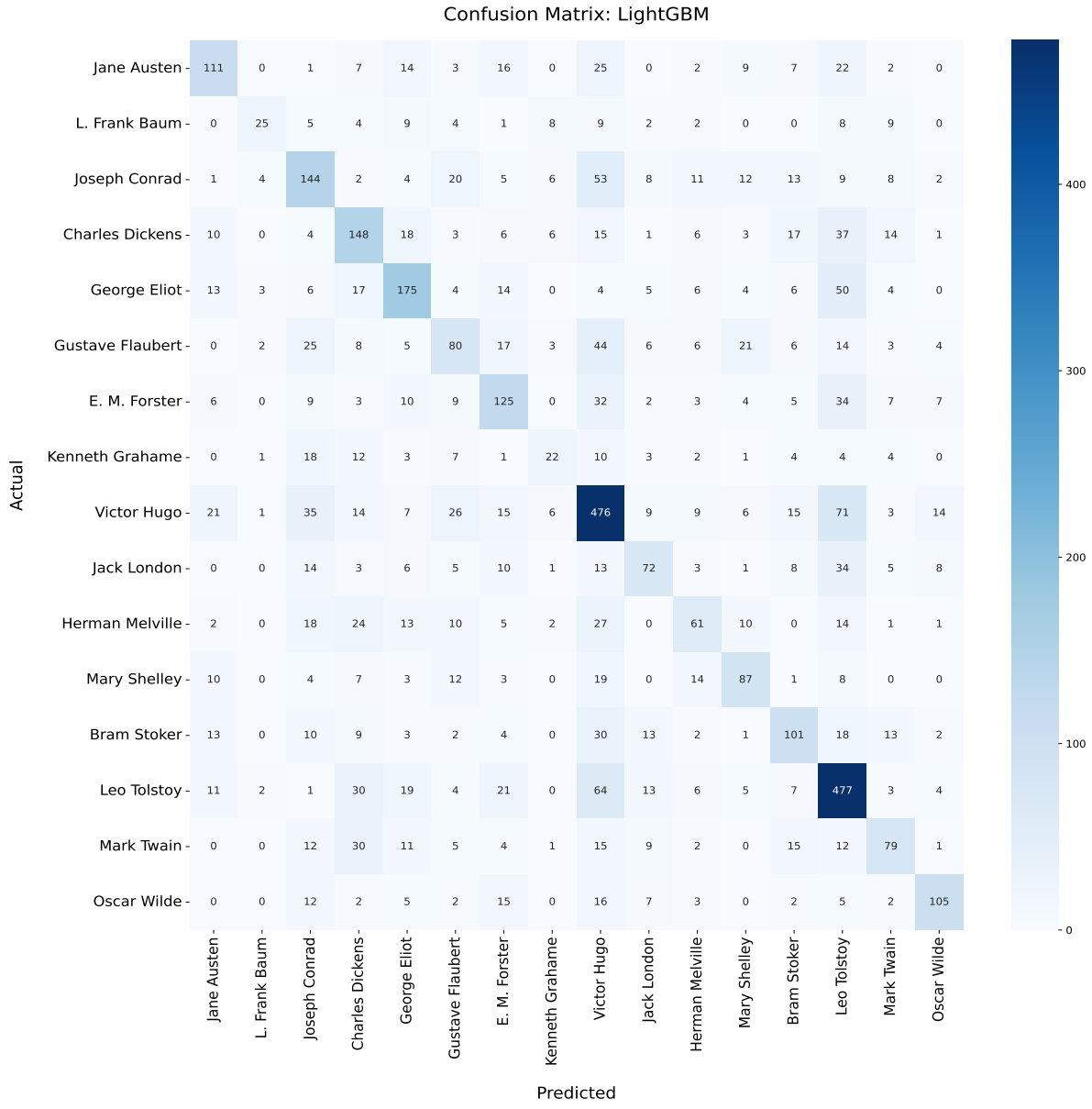


Figure 4.7: Accuracy Confusion Matrix *LightGBM*: (16 Authors).

Figure 4.8 is an Accuracy Matrix belonging to the multiclass classification task of the 4 authors that displays the model *Accuracy* performance per author. Evidently, in this task, the model performs well on the authors *Victor Hugo* and *Leo Tolstoy* that hold the top text segments among all authors compared to the other two authors in this task.

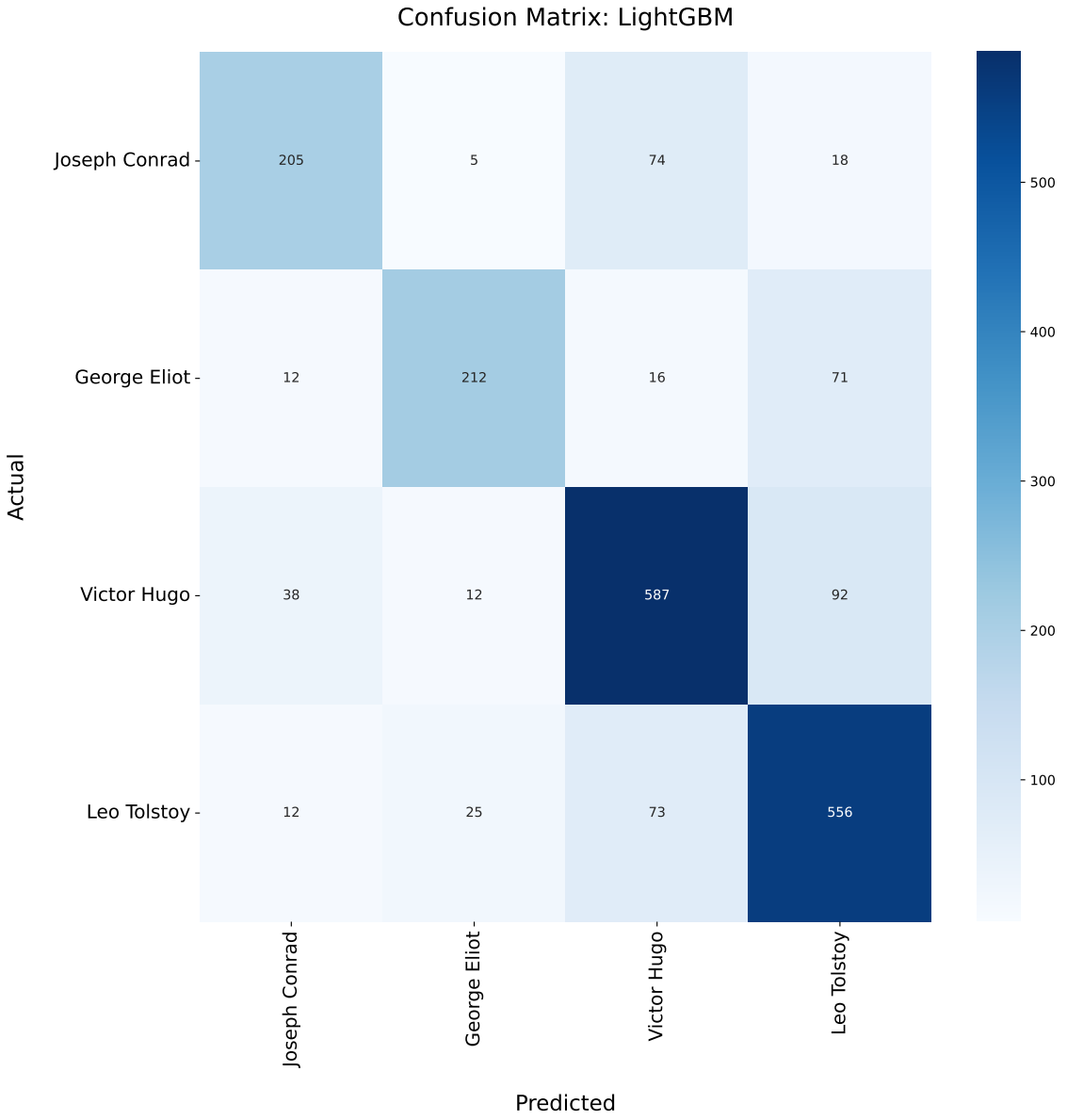
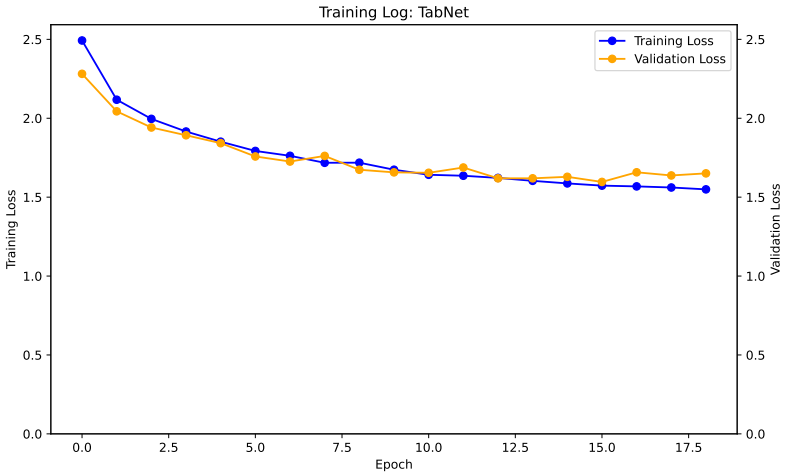


Figure 4.8: Accuracy Confusion Matrix *LightGBM*: (4 Authors).

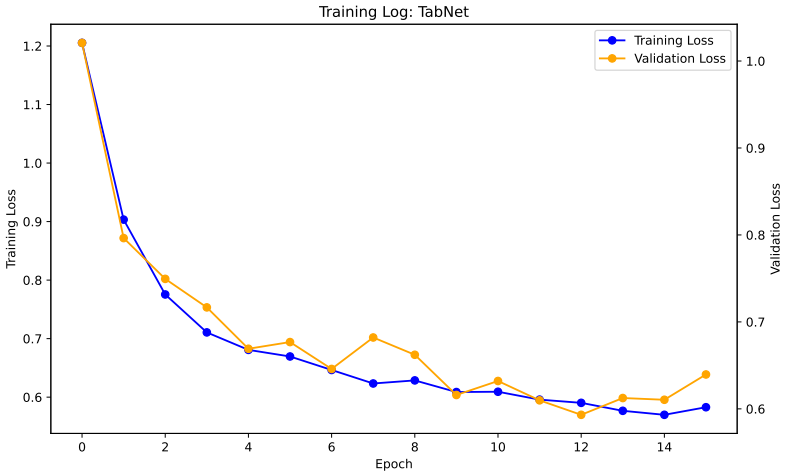
4.3.2 TabNet

For the training progress, Figure 4.9a and 4.9b show the model's performance for both tasks. In the multiclass classification task of 16 authors, the *Training Loss* starts at around 2.5 and consistently decreases slowly across *epochs* to about 1.5. However, the *Validation Loss* begins at approximately 2.3 and increases to around 1.6 over the *epochs*. The model had an early stopping at *epoch* 18, choosing *epoch* 15, and the best loss margin was around 1.59647, based on the validation set.

As for the multiclass classification task of 16 authors, the *Training Loss* starts at around 1.2 and decreases slowly across *epochs* to about 0.6. For the *Validation Loss*, it begins at approximately 1.2 also and decreases to around 0.65 over the *epochs*. The model had an early stopping at *epoch* 15, choosing *epoch* 12, and the best loss margin was around 0.59316, based on the validation set.



(a) (16 Authors).



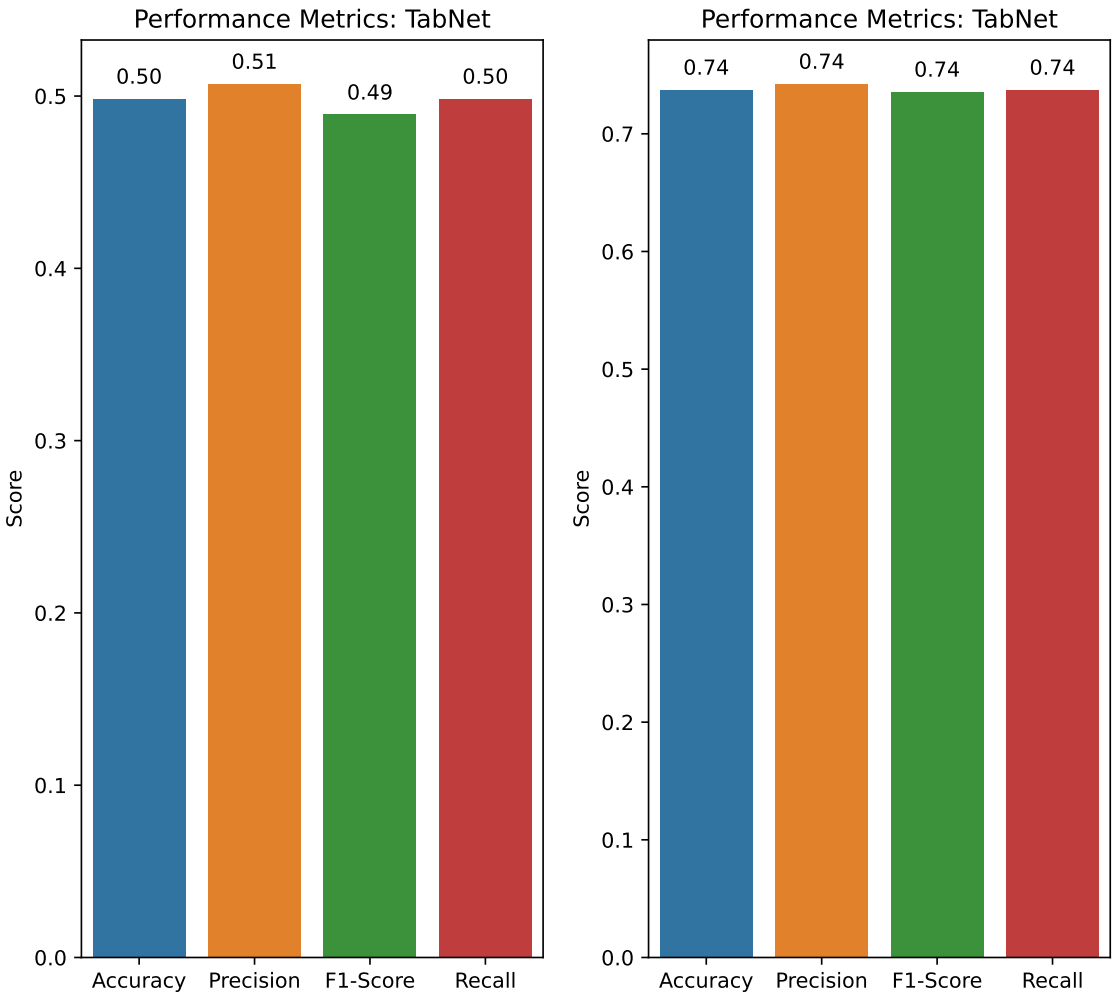
(b) (4 Authors).

Figure 4.9: Training Progress: *TabNet*.

Regarding the performance for both tasks, Figure 4.10a and 4.10b show the scores according to the metrics. In the multiclass classification task of 16 authors, the model achieves an *Accuracy* of around 50%, implying that it correctly predicts the authors about half the time. The *Accuracy* is not optimal and is in the same proximity as the multiclass classification task of 16 authors in Section 4.3.1.

For the multiclass classification task of 4 authors, the model achieves an *Accuracy* of around 74%, establishing reasonable performance. Similarly, the model's *Accuracy* in the multiclass classification task of 4 authors lies within the same proximity as that of 4 authors in Section 4.3.1.

Because of the architecture of *TabNet*, which is *Neural Network* architecture, it may not work well on imbalanced data or small data [41][40]. Consequently, a larger dataset may be required to handle the multiclass classification of 16 authors better.



(a) (16 Authors).

(b) (4 Authors).

Figure 4.10: Performance Metrics: *TabNet*.

Figure 4.11 is an Accuracy Matrix belonging to the multiclass classification task of the 16 authors that displays the model *Accuracy* performance per author. The model *Accuracy* shows a more or less similar manner to the results of the task involving *LightGBM* in the multiclass classification task of the 16 authors, in Section 4.3.1; the model performed better on the authors with a high number of text segments than those with a lower number of text segments.

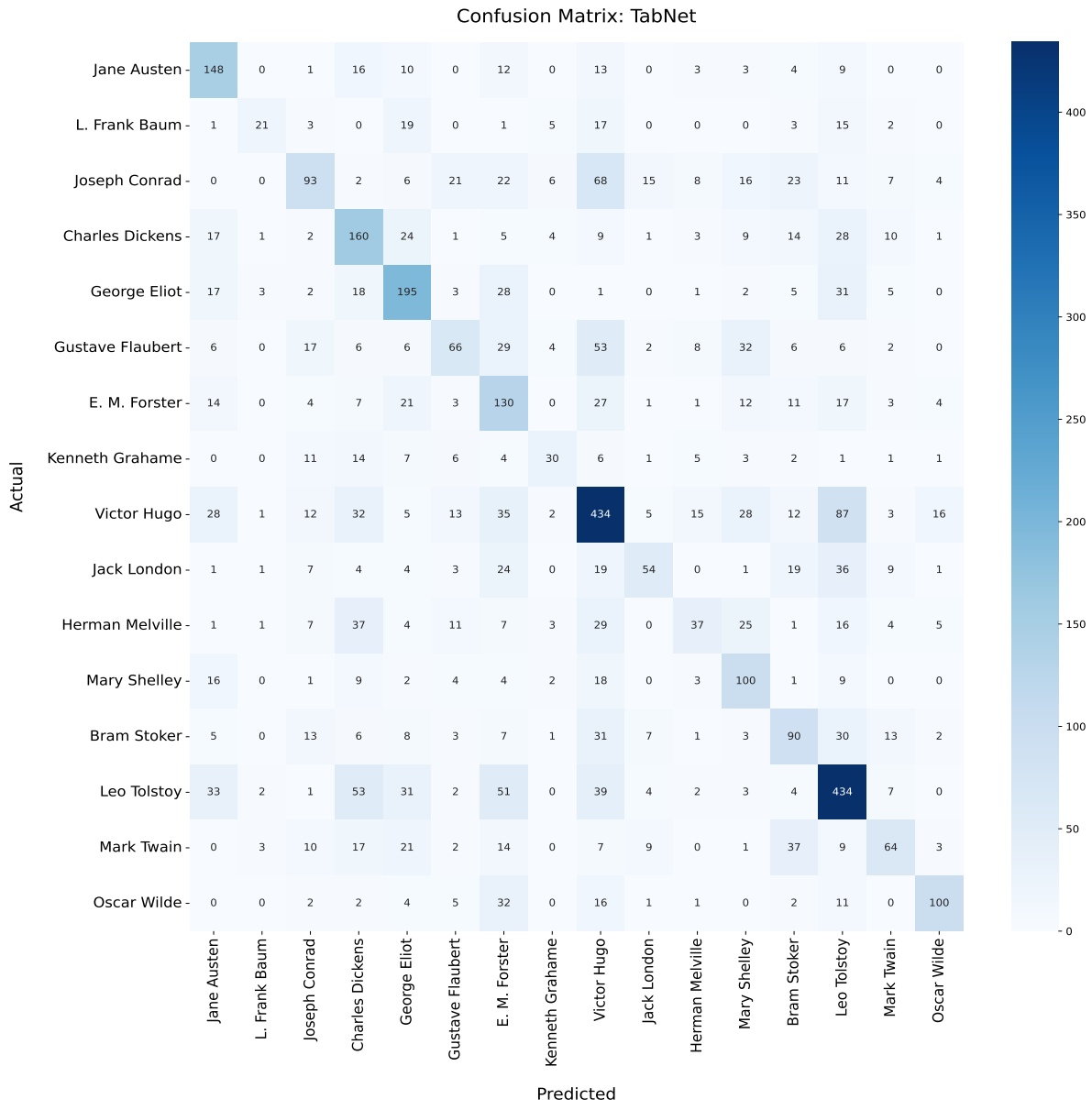


Figure 4.11: Accuracy Confusion Matrix *TabNet*: (16 Authors).

Figure 4.12 is an Accuracy Matrix belonging to the multiclass classification task of the 4 authors that displays the model *Accuracy* performance per author. The *Accuracy* of the model in this task also more or less is similar to the results of the task involving *LightGBM* in the multiclass classification task of the 4 authors, in Section 4.3.1; the model performs well on the authors *Victor Hugo* and *Leo Tolstoy* that hold the top text segments among all authors compared to the other two authors in this task.

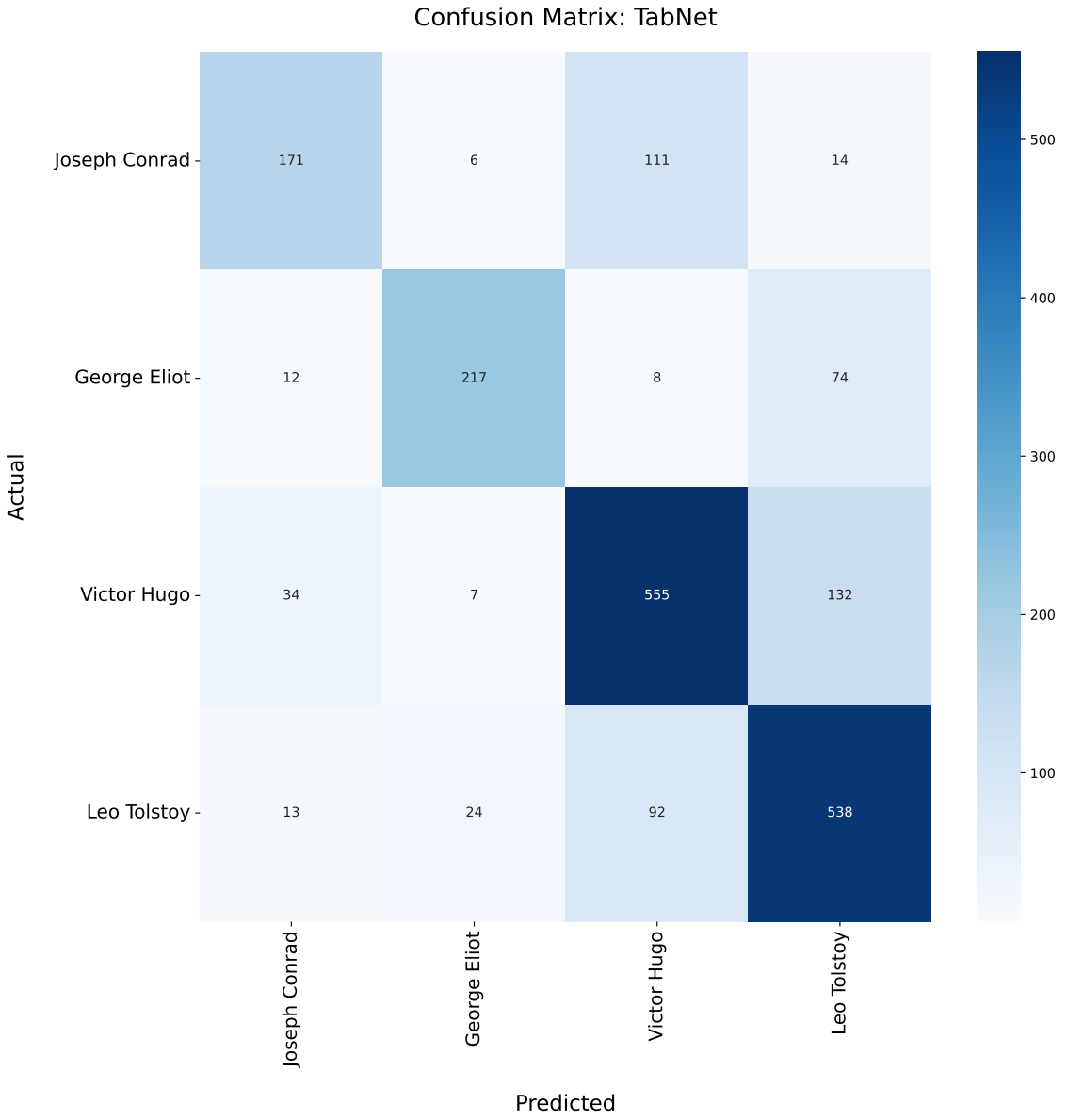


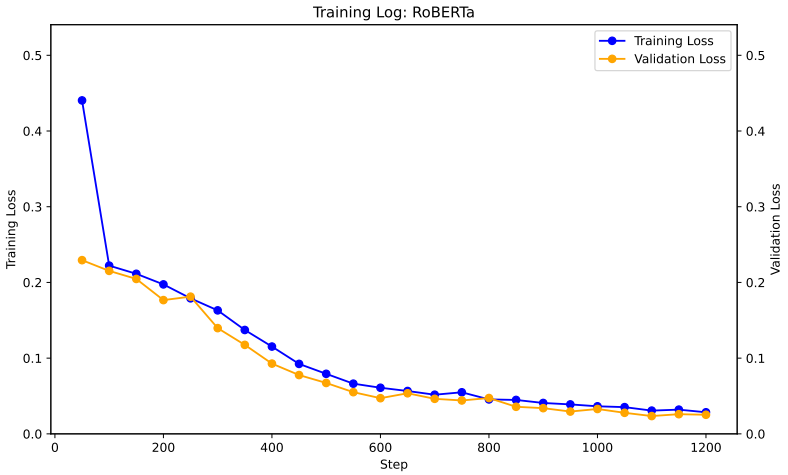
Figure 4.12: Accuracy Confusion Matrix *TabNet*: (4 Authors).

4.4 RQ3

This section presents the result related to **RQ3**. For the model's training progress in both tasks, Figure 4.13a and 4.13b show the progress for both tasks.

In the multiclass classification task of 16 authors, the *Training Loss* starts at around 0.45 but drops rapidly to around 0.25 in the early steps, then declines steadily and plateaus around 0.02. The *Validation Loss* reflects the same manner as the *Training Loss*, begins from around 0.225, and decreases consistently but not at the same rate. The model stopped at *step* 1200, choosing *step* 1100, and the best evaluation loss margin was around 0.023516, based on the validation set.

When it comes to the task multiclass classification task of 4 authors, the *Training Loss* starts from around 0.55, decreases rapidly, and stabilizes around *step* 150 and plateaus around 0.05 in *step* 300. For the *Validation Loss*, it starts from around 0.45, decreases rapidly, and plateaus around 0.06. The model stopped at *step* 300, choosing *step* 250, and the best evaluation loss margin was around 0.069087, based on the validation set.

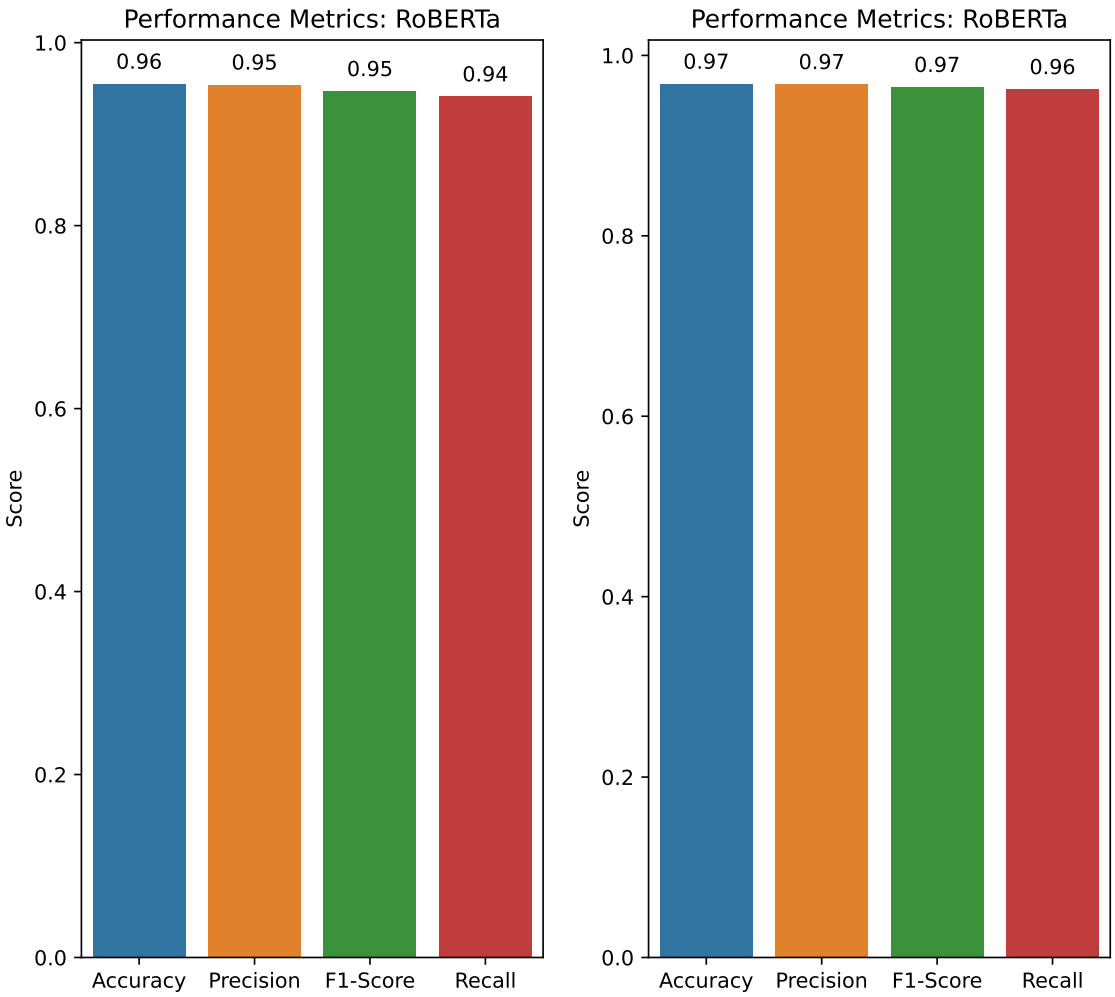


(a) (16 Authors).

(b) (4 Authors).

Figure 4.13: Training Progress: *RoBERTa*.

Regarding the performance for both tasks, Figure 4.14a and 4.14b show the scores according to the metrics. In the multiclass classification task of 16 authors, the model achieves an *Accuracy* of around 96%, implying a robust performance. For the multiclass classification task of 4 authors, the model achieves an *Accuracy* of around 97%, establishing high performance. Additionally, the scores of *Precision*, *F1-score*, and *Recall* in both tasks are in a similar range, suggesting that there is a balanced performance.



(a) (16 Authors).

(b) (4 Authors).

Figure 4.14: Performance Metrics: *RoBERTa*.

It is worth noting that according to the results of Section 4.1 that **Cluster 2** had the longest sentences on average, 31.78 tokens per sentence, indicating that making the segment size 15-sentence per segment was reasonable because it did not exceed the limits of *RoBERTa*. Thus, the total number of tokens would be as follows:

$$31.78 \times 15 = 476.7 < 512$$

Figure 4.15 is an Accuracy Matrix that displays the model *Accuracy* performance per Author.

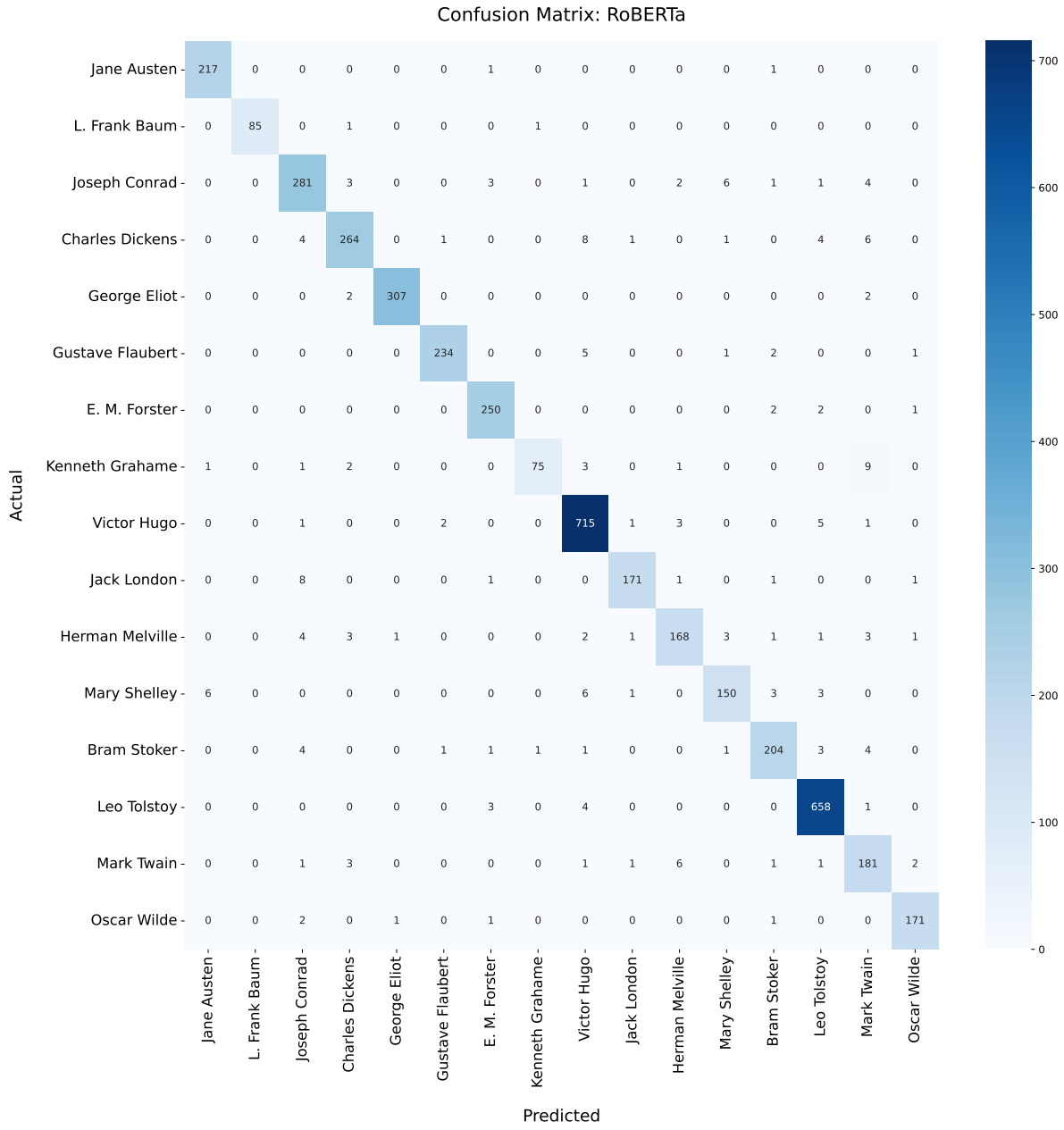


Figure 4.15: Accuracy Confusion Matrix *RoBERTa*: (16 Authors).

Figure 4.16 is an Accuracy Matrix that displays the model *Accuracy* performance per Author.

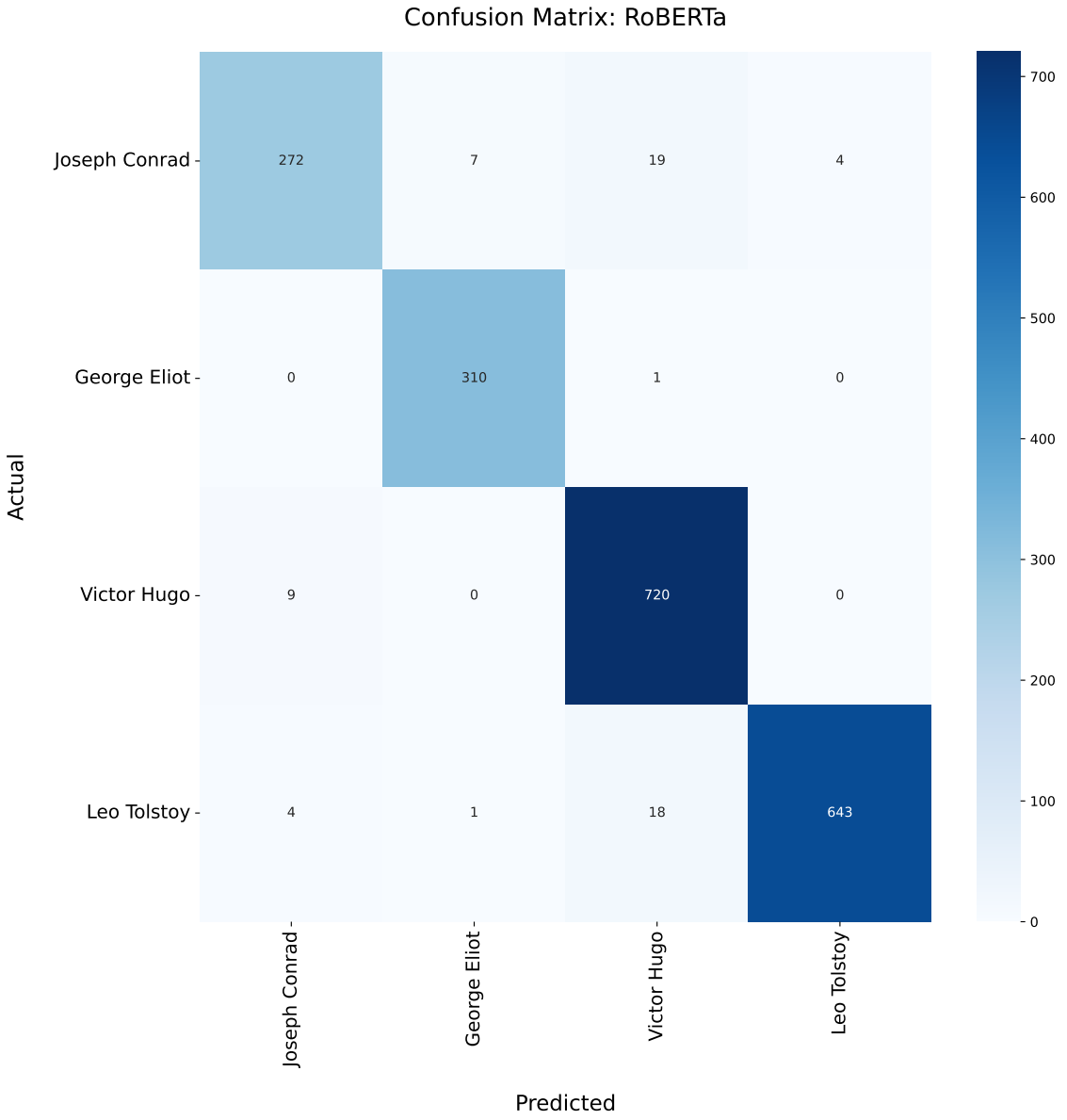


Figure 4.16: Accuracy Confusion Matrix *RoBERTa*: (4 Authors).

4.5 RQ4

According to the results of Section 4.1, there is an imbalance among the classes in the dataset. Nevertheless, the result of Section 4.4 demonstrates that *RoBERTa* has shown high performance in both classification tasks, meaning it has handled the imbalance gracefully, and the size and imbalance of the dataset did not impact its performance and achieved high performance. It has achieved high performance in the authorship attribution because it is a pretrained model.

5 Related Work

This section presents related works and their descriptions and compares their results with the results of this study. The research in the analysis of the computer-generated literature is extensive. After thorough research, this project has considered some studies that resonated more with the objectives of this study:

- H. Elahi and H. Muneer, *Identifying different writing styles in a document intrinsically using stylometric analysis* [11].
- L. Yang, G. Wang, and H. Wang, *Reimagining literary analysis: Utilizing artificial intelligence to classify modernist french poetry* [22].
- H. O. Hatzel, H. Stiemer, C. Biemann, and E. Gius, *Machine learning in computational literary studies* [44].
- D. K. Wendt, *Recognizing literary merit with deep learning* [45].
- T. Schmidt, K. Dennerlein, and C. Wolff, *Emotion classification in German plays with transformer-based Language Models pretrained on historical and contemporary language* [46].
- Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, Omer Levy; *SpanBERT: Improving Pre-training by Representing and Predicting Spans* [47].
- M. Parigini and M. Kestemont, *The roots of doubt. Fine-Tuning a BERT model to explore a stylistic phenomenon* [48].
- X. Zhang, F. Wei, and M. Zhou, *HIBERT: document level pre-training of hierarchical bidirectional transformers for document summarization* [49].
- Qian Li, Hao Peng, Jianxin Li, Congying Xia, Renyu Yang, Lichao Sun, Philip S. Yu, and Lifang He. *A Survey on Text Classification: From Traditional to Deep Learning* [50].
- Elmahdy, A., Inan, H.A., Sim, R. *Privacy Leakage in Text Classification: A Data Extraction Approach* [51].

5.1 Description Of The Studies

H. Elahi and H. Muneer, *Identifying different writing styles in a document intrinsically using stylometric analysis*. In 2018, *H. Elahi* and *H. Muneer* developed a solution that can classify various writing styles within a document through stylometric analysis by dividing the document into partitions and using different metrics to compute lexical, vocabulary richness, readability, and text structure features and vectorizing the resulting features to group the writing styles according to the similarities using unsupervised machine learning [11]. The study of this project adopted a similar approach to the study of *H. Elahi* and *H. Muneer*. It segmented the book texts and performed stylometric analysis to compute the textual features of the text segments by applying the same stylometric metrics criteria used in the work of *H. Elahi* and *H. Muneer*.

L. Yang, G. Wang, and H. Wang, *Reimagining literary analysis: Utilizing artificial intelligence to classify modernist French poetry*. The authors of the study applied feature

extraction techniques such as *TF-IDF* and *Doc2Vec* and machine learning algorithms such as *SVM*—the model in the study aimed to classify poems by their stylistic and thematic attributes. The methodology of the latter work started with text preprocessing for categorization and vectorization, followed by Feature Extraction *TF-IDF* and *Doc2Vec*. The comparison of classification algorithms has compared, adjusted and optimized the *SVM* parameters and investigated other classifiers such as *Logistic Regression (LR)*, *Bagging*, *Random Forest (RF)*, *AdaBoost*, *Gradient Boosted Decision Trees (GBDT)*, *XGBoost*, and *LightGBM* using metrics such as *Accuracy*, *F1 score*, *Precision* and *Recall*. The three significant contributions of the study focused on the following: Innovating literary research with *AI* by using *SVM*, *TF-IDF*, and *Doc2Vec*, enhancing Literary Comprehension by analyzing stylistic elements and opening new possibilities for literary discovery and Technical Enrichment of *AI* by developing more efficient methods of analysis [22].

H. O. Hatzel, H. Stiemer, C. Biemann, and E. Gius, *Machine learning in computational literary studies*. Apart from the earlier related works, this work aided this project in defining its boundaries and gaining an overview of the machine learning field applied to computational literary studies. The work surveyed the machine learning methodologies employed by numerous scientific publications and compared their scope and effectiveness. It screened 215 papers and focused on 40 to understand the machine learning methods and the types of the models. This research helped this project understand the approaches available in the field and informed its decision-making practice. For instance, the *Transformer* type is the most popular method employed in the field, followed by *SVM* and *Logistic Regression* by some distance. The examination's findings shed light on the dynamic evolution of the *NLP* community's approach. It revealed a gradual shift away from the traditional pipeline approach towards more advanced end-to-end models like the *Transformer* models. This shift underscores the potential for future advancements in the field. It has also shown how recent advancements in machine learning techniques are reducing processing times and allowing the processing of longer texts while giving rise to new opportunities for automation. Furthermore, it also highlighted that the traditional pipeline-based approach still complements modern methods [44].

D. K. Wendt, *Recognizing literary merit with deep learning*. Regarding *Deep Learning*, this project trained a *Deep Learning* model to recognize literary merit in English literature. As the first step, it used a *distilBERT* model on a corpus of 30 novels and achieved a final test set *Accuracy* of 92%. The second step was to extend the corpus to a more extensive set of 108 books, on which it achieved an *Accuracy* of 75%. It used the same source as this study, the free public domain *Project Gutenberg*. It employed *distilBERT* for classification, and the model can classify sequences of up to 512 tokens. The text files for each book have been preprocessed by tokenizing the entire texts into sequences of 512 tokens, which are then independently run via *distilBERT*. The model learns a distilled or approximate version of *BERT*, meaning it suffers a 5% degradation from *BERT* and retains 95% of the performance but uses only half of the parameters of *BERT* [45].

T. Schmidt, K. Dennerlein, and C. Wolff, *Emotion classification in German plays with transformer-based Language Models pre-trained on historical and contemporary language*. This study presents the results of classifying emotions in historical German plays. It conducted textual annotation of 11 plays and acquired 13000 emotion annotations. They evaluated multiple traditional machine-learning approaches with *Transformer*-based models. It achieved classification *Accuracy* of up to 90%, although lower *Accuracy* occurred in settings with more classes. The evaluated models are *Transformer*-based language models like *BERT* and *ELECTRA* [46].

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, Omer Levy, *SpanBERT: Improving Pre-training by Representing and Predicting Spans*. The study used a pretraining method called *SpanBERT*, designed to represent and predict text spans. It offered better *Accuracy* than *BERT*. However, the authors used a different random process to mask token spans rather than individual ones and by sampling a single contiguous text segment for each training example instead of two [47].

M. Parigini and M. Kestemont, *The roots of doubt. Fine-Tuning a BERT model to explore a stylistic phenomenon*. Another study explores fine-tuning and customization of the *BERT* model, but this time explores the stylistic characteristics of Italian author *Italo Calvino*. The study aims to model dubitative text's presence in work [48].

X. Zhang, F. Wei, and M. Zhou, *HIBERT: document level pre-training of hierarchical bidirectional transformers for document summarization*. *HIBERT* stands for *HIerarchical Bidirectional Encoder Representations* from *Transformers*. The study has used it for document encoding and pretraining of the document using unlabeled data. The work aimed to automate document summarization by rewriting its content into a shorter form while preserving it [49].

Qian Li, Hao Peng, Jianxin Li, Congying Xia, Renyu Yang, Lichao Sun, Philip S. Yu, and Lifang He, *A Survey on Text Classification: From Traditional to Deep Learning*. This study surveys the methods, metrics, and datasets for text classification. It reviews the *State-Of-The-Art* approaches to text classification from 1961. It presents a detailed comparison of the results obtained in many studies based on different models, metrics, datasets, and types of classification [50].

Elmahdy, A., Inan, H.A., Sim, R., *Privacy Leakage in Text Classification: A Data Extraction Approach*. This study highlights a significantly overlooked issue of immense weight in privacy and ethics in text classification. It addresses the potential for security and privacy leakage and the memorization of training data. Furthermore, it highlights the importance of auditing strategies to assess any potential security issue with personal data used without consent [51].

5.2 Comparison And Discussion

When it comes to the work of *H. Elahi* and *H. Muneer*, this project adopted the approach of the latter work. Still, this study differs moderately from the latter work in two different areas; this study aims to group the writing styles to analyze the linguistic features of the classic literature to understand their properties and to use the textual features of the text segments for classification tasks to asses the application of this approach in authorship attribution. Hence, while the work of *H. Elahi* and *H. Muneer* used this approach and focused merely on unsupervised techniques and grouping writing styles; this study used not only unsupervised techniques for analyzing the linguistic features of classic literature but instead extended it to advanced and modern supervised machine algorithms also for a further task that is, authorship attribution [11].

Regarding the study of poem classifications, it successfully classified poetry with an *Accuracy* of 74.3%. Still, it also showed the limitations of a traditional approach, its limitations in capturing features, and its generalization ability. Furthermore, it used a small dataset. With *SVM* using ensemble features, it provided the best performance of 74.3%, with *LightGBM* reaching the second best *Accuracy* of 73.7%. Similarly, in this study, *LightGBM* was the second best performer after *RoBERTa* with an *Accuracy* of 78% on a more diverse dataset. This study applied a methodology similar to text preprocessing, feature extraction, and algorithm comparison to the related work. This project tried to

build up on the limitations illustrated by the related work in question, which led to employing the *State-Of-The-Art* techniques that would offer a better capability to capture nuances more effectively. The result is that this project's modern approach has produced much better *Accuracy* and a more detailed analysis of the dataset, overcoming some of the limitations of the related study in question [22].

While this project initially considered a traditional pipeline approach, the study *Machine learning in computational literary studies* aided in understanding that the *NLP* community has been moving away from this approach and towards an end-to-end model such as the *Transformer* models. Thus, the latter has oriented the decision-making process towards adopting a *Transformer* model. The related study's statistics revealed that *Transformer* models demonstrated superior performance. However, they also presented challenges that this project encountered, such as the demand for computational resources and the scarcity of training data. *Transformer* models typically require training with billions of tokens, as exemplified by the original *BERT* implementation's use of a corpus exceeding three billion words. The project's findings provided compelling evidence that *Transformer* models consistently outperform more traditional approaches, even without specific training data. This robust validation strongly justified this project's adoption of a *Transformer* model. Most performance metrics employed are also prevalent in the machine learning community, as illustrated by the study [44].

Additionally, the work of Mr. *Wendt* was used to understand the *Transformer* approach when analyzing a corpus of English novels, specifically from *Project Gutenberg*'s accessible repository of public domain works. The results of this study obtained a test set scoring of over 98.4% for *Accuracy*. The comparison present in this study between *distilBERT* and other versions, such as *RoBERTa* and *BERT*, was used to guide this project's learning curve. Hence, the comparison facilitated the final decision, which favored the robustly *Optimized BERT Approach*, known as *RoBERTa*, which was introduced by *Meta* and offered a significant performance improvement over *distilBERT* and *BERT*. *distilBERT* has a faster inference speed but compromises speed over prediction *Accuracy*. *RoBERTa*, instead, has the best prediction metrics but is computationally intensive and slower. As usual, a trade-off between computation and prediction metrics must be carefully considered at the start of the project [45].

The related study of emotion classification in German plays further corroborated this project's final choice to proceed with a *Transformer*-based model. Furthermore, it informed this project on the necessity of being careful when it comes to the trap of generalization. It is important to note that relying solely on dramatic texts for training could lead to generalization issues. This cautionary note underscores the need for a balanced and diverse training dataset. The related study showed no significant improvement even if *Transformer*-based models were trained on relevant historical data; the study plans to extend exploration by acquiring large amounts of general text material (not dramatic) for the specific time frame to check if performance improvement is possible. A similar approach would also be interesting for this project [46].

This project considered possible customization of models to increase performance and improve the understanding of specific aspects of the novels. The following three studies helped break the black box of *Transformer* models such as *BERT* and obtain information for possible customizations.

The *SpanBert* study showed how it is possible to customize and implement new methods. Although this work did not employ such levels of customization, it was crucial to understand the inner workings of *BERT* and consider possible alternatives for future works [47]. The roots of Doubt instead provided information for fine-tuning a *BERT* model to

capture stylistic features that are not immediately evident [48].

The *HIBERT* study was another exciting innovation that showed how to pretrain document-level hierarchical bidirectional transformer encoders on unlabeled data. It showed the potential of applying similar models to tasks that required hierarchical encoding and was another example of an innovative approach to finding better and new methods for specific challenges [49].

A Survey on text classification is one of the studies that inspired this project to consider using *XGBoost* first and then *LightGBM*, considering the outstanding performance demonstrated here. *RoBERTa* also came on top regarding performance and *Accuracy* on various tasks. The most engaging part of the related study was discussing future research challenges. It articulately explains the different challenges affecting datasets (zero-shot, few-shot, special domains), models (text representation, model integration, and efficiency), and performance (semantic robustness, interpretability), all thoroughly explained and analyzed. Its summary tables also allowed a quick and effective comparison of the different models and metrics in the context of different applications for text classification, from sentiment analysis to text classification to topic labeling and more [50].

Another vital study is on privacy leakage in text classification. It is an often overlooked aspect worth considering. Even if the public domain dataset does not present privacy issues, some of the models a work could use in future work could contain personal data, and this project considered it was worth being aware of the security and privacy implications. Memorization of extensive training data can have unintended consequences. Training data can sometimes include private data that may be inadvertently leaked. The work explores the privacy and security implications of performing text classification on sensitive or private data, and it is a starting point for a broader discussion on such a vital matter [51].

6 Conclusion

The study suggests that segmenting text and quantifying it using stylometry analysis, whether for analyzing text or conducting authorship attribution, is feasible and worth discovering. Regarding literature analysis, the study uncovered the books' characteristics reasonably well, offering insights on what to consider regarding linguistic features to obtain a complex yet engaging and easy-to-read text.

For authorship attribution using the quantified values of the textual features, the approach yielded results indicating that this approach applies to such tasks. The approach was utilized on the downsized dataset of 4 authors, and the performance enhanced, indicating class imbalance may have impacted the overall performance. Nonetheless, this approach still needs improvements, such as filtering the stylometrics features to work on smaller metrics that yield the most information gain and perform better. Regarding *RoBERTa*, it has achieved high performance regardless of the data imbalance, showing robustness in this task.

Undoubtedly, the biggest hurdle the study faced was the limitation of computing power and memory, which significantly slowed the progress. It did not allow for enhancement of the number of books per author to extend the dataset to evaluate the approach using a larger dataset or pursue optimization processes.

Lastly, this project also conducted additional work in the Appendix, Section A.2, merely as a side contribution to this study by performing an exploratory analysis. The exploratory analysis involves utilizing *RoBERTa* to tokenize and embed the text segments and using the resulting embedding to train *Latent Dirichlet Allocation (LDA)* for topic modeling. This objective is merely for further exploratory experiments.

7 Future Work

In terms of further work, the study recommends replicating the methodology of this study by considering and improving multiple aspects. The first aspect is to analyze the stylometry metrics to obtain only a smaller set yet more focused on different and unique linguistic features. In this study, there were metrics that more or less provided the same information, and using them together may not be optimal or provide significant information gain; on the contrary, they may impact the learning process, make the dataset highly dimensional unnecessarily, and may make the learning process demanding in computational resources.

The second aspect is enhancing the number of books per author, preferably by applying it to authors with large arrays of books to generate a large dataset. Another aspect to consider is adjusting the segment size to different lengths, meaning testing different sizes to determine the optimal size that may lead to the most significant information gain and higher performance.

Lastly, another recommendation is to use this approach with books from different eras to evaluate how writing styles have evolved and what linguistic features have stayed consistent over time to provide novel insights about using language as a tool for human expression in written form.

8 Acknowledgement

In this thesis project, the *Generative Tools* used during the study are *Grammarly* for grammar checks, auto-correction, and text refinement, *Google Scholar* for literature review and retrieving relevant references regarding *Mathematics*, *NLP*, *Text Mining*, *Machine Learning* algorithms and *AI* models; *TabNet*, *LightGBM*, *BERT* and *RoBERTa*. In addition, the study used *ChatGPT* to construct the layout of the *LaTeX* document used for the report of this study.

9 Contributions

The authors of this project contributed to this study with more or less the same degree of contributions, where both engaged incrementally in each phase of the process together.

References

[1] A. V. Miller, *Pairing young adult and classic literature in the high school English curriculum*. The University of Maine, 2017.

[2] TES, "Oliver twist: Full scheme of work," 2018, accessed: 2023-04-14. [Online]. Available: https://www.tes.com/teaching-resource/oliver-twist-full-scheme-of-work-11525438

[3] E. Grant, "The importance of classical literature in secondary education," *Canyon Journal of Undergraduate Research*, vol. 1, no. 1, 2023.

[4] Y. Zhao and J. Zobel, "Searching with style: Authorship attribution in classic literature," in *Proceedings of the thirtieth Australasian conference on Computer science-Volume 62*. Citeseer, 2007, pp. 59–68.

[5] H. Dry, "Syntax and point of view in jane austen's" emma"," *Studies in Romanticism*, pp. 87–99, 1977.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[7] P. Johri, S. K. Khatri, A. T. Al-Taani, M. Sabharwal, S. Suvanov, and A. Kumar, "Natural language processing: History, evolution, application, and future work," in *Proceedings of 3rd International Conference on Computing Informatics and Networks: ICCIN 2020*. Springer, 2021, pp. 365–375.

[8] D. Rothman, *Transformers for Natural Language Processing: Build innovative deep neural network architectures for NLP with Python, PyTorch, TensorFlow, BERT, RoBERTa, and more*. Packt Publishing Ltd, 2021.

[9] T. Teubner, C. M. Flath, C. Weinhardt, W. van der Aalst, and O. Hinz, "Welcome to the era of chatgpt et al. the prospects of large language models," *Business & Information Systems Engineering*, vol. 65, no. 2, pp. 95–101, 2023.

[10] J. Savoy, "Machine learning methods for stylometry," *Cham: Springer*, 2020.

[11] H. Elahi and H. Muneer, "Identifying different writing styles in a document intrinsically using stylometric analysis," *The complete code and detailed documentation is available on the attached Github Link: https://github. com/harismuneer/Writing-Styles-Classification-Using-Stylometric-Analysis*, 2018.

[12] M. H. Freeman, "Cognitive linguistic approaches to literary studies: State of the art in cognitive poetics," 2009.

[13] H. Zhao, H. Chen, F. Yang, N. Liu, H. Deng, H. Cai, S. Wang, D. Yin, and M. Du, "Explainability for large language models: A survey," *ACM Transactions on Intelligent Systems and Technology*, vol. 15, no. 2, pp. 1–38, 2024.

[14] P. Ganesh, Y. Chen, X. Lou, M. A. Khan, Y. Yang, H. Sajjad, P. Nakov, D. Chen, and M. Winslett, "Compressing large-scale transformer-based models: A case study on bert," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 1061–1080, 2021.

[15] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[16] Q. Cao, Y. K. Lal, H. Trivedi, A. Balasubramanian, and N. Balasubramanian, "Irene: Interpretable energy prediction for transformers," *arXiv preprint arXiv:2106.01199*, 2021.

[17] S. Rebora, "A digital edition between stylometry and ocr," *Textual Cultures*, vol. 12, no. 2, pp. 71–90, 2019.

[18] P. Baxi, "Sentiment analysis based on social networks using support expectation-maximization for e-commerce applications," in *2023 International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE)*. IEEE, 2023, pp. 1–6.

[19] P. Dewan, A. Kashyap, and P. Kumaraguru, "Analyzing social and stylometric features to identify spear phishing emails," in *2014 apwg symposium on electronic crime research (ecrime)*. IEEE, 2014, pp. 1–13.

[20] A. Rocha, W. J. Scheirer, C. W. Forstall, T. Cavalcante, A. Theophilo, B. Shen, A. R. Carvalho, and E. Stamatatos, "Authorship attribution for social media forensics," *IEEE transactions on information forensics and security*, vol. 12, no. 1, pp. 5–33, 2016.

[21] I. Rep and V. Čeperić, "Boosting the performance of transformer architectures for semantic textual similarity," *arXiv preprint arXiv:2306.00708*, 2023.

[22] L. Yang, G. Wang, and H. Wang, "Reimagining literary analysis: Utilizing artificial intelligence to classify modernist french poetry," *Information*, vol. 15, no. 2, 2024. [Online]. Available: https://www.mdpi.com/2078-2489/15/2/70

[23] D. Kopev, D. Zlatkova, K. Mitov, A. Atanasov, M. Hardalov, I. Koychev, and P. Nakov, "Recursive style breach detection with multifaceted ensemble learning," in *Artificial Intelligence: Methodology, Systems, and Applications: 18th International Conference, AIMSA 2018, Varna, Bulgaria, September 12–14, 2018, Proceedings 18*. Springer, 2018, pp. 126–137.

[24] S. Jacques, "Machine learning methods for stylometry," 2020.

[25] J. P. Kincaid, R. P. Fishburne Jr, R. L. Rogers, and B. S. Chissom, "Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel," 1975.

[26] R. Flesch, "A new readability yardstick." *Journal of applied psychology*, vol. 32, no. 3, p. 221, 1948.

[27] A. Khodabakhsh, F. Yesil, E. Guner, and C. Demiroglu, "Evaluation of linguistic and prosodic features for detection of alzheimer's disease in turkish conversational speech," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, pp. 1–15, 2015.

[28] G. Stone and L. D. Parker, "Developing the flesch reading ease formula for the contemporary accounting communications landscape," *Qualitative Research in Accounting & Management*, vol. 10, no. 1, pp. 31–59, 2013.

[29] D. Eleyan, A. Othman, and A. Eleyan, "Enhancing software comments readability using flesch reading ease score," *Information*, vol. 11, no. 9, p. 430, 2020.

[30] J. B. Kouame, "Using readability tests to improve the accuracy of evaluation documents intended for low-literate participants," *Journal of MultiDisciplinary Evaluation*, vol. 6, no. 14, pp. 132–139, 2010.

[31] G. R. Klare, "Assessing readability," *Reading research quarterly*, pp. 62–102, 1974.

[32] D. Pavelec, E. Justino, and L. S. Oliveira, "Author identification using stylometric features," *Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial*, vol. 11, no. 36, pp. 59–65, 2007.

[33] R. H. Baayen, *Word frequency distributions*. Springer Science & Business Media, 2001, vol. 18.

[34] E. R. Dougherty, *Probability and statistics for the engineering, computing, and physical sciences*. Prentice-Hall, Inc., 1990.

[35] S. Kappal *et al.*, "Data normalization using median median absolute deviation mmad based z-score for robust predictions vs. min–max normalization," *Lond. J. Res. Sci. Nat. Form*, vol. 19, no. 10.13140, 2019.

[36] I. Cohen, Y. Huang, J. Chen, J. Benesty, J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," *Noise reduction in speech processing*, pp. 1–4, 2009.

[37] J. Feng, Y. Yu, and Z.-H. Zhou, "Multi-layered gradient boosting decision trees," *Advances in neural information processing systems*, vol. 31, 2018.

[38] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," *Advances in neural information processing systems*, vol. 30, 2017.

[39] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, "A comparative analysis of gradient boosting algorithms," *Artificial Intelligence Review*, vol. 54, pp. 1937–1967, 2021.

[40] S. Ö. Arik and T. Pfister, "Tabnet: Attentive interpretable tabular learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 8, 2021, pp. 6679–6687.

[41] K. McDonnell, F. Murphy, B. Sheehan, L. Masello, and G. Castignani, "Deep learning in insurance: Accuracy and model interpretability using tabnet," *Expert Systems with Applications*, vol. 217, p. 119543, 2023.

[42] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[43] K. Clark, U. Khandelwal, O. Levy, and C. D. Manning, "What does bert look at? an analysis of bert's attention," *arXiv preprint arXiv:1906.04341*, 2019.

[44] H. O. Hatzel, H. Stiemer, C. Biemann, and E. Gius, "Machine learning in computational literary studies," *it - Information Technology*, vol. 65, no. 4-5, 2023. [Online]. Available: https://doi.org/10.1515/itit-2023-0041

[45] D. K. Wendt, "Recognizing literary merit with deep learning," 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:236470121

[46] T. Schmidt, K. Dennerlein, and C. Wolff, "Emotion classification in German plays with transformer-based language models pretrained on historical and contemporary language," in *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, S. Degaetano-Ortlieb, A. Kazantseva, N. Reiter, and S. Szpakowicz, Eds. Punta Cana, Dominican Republic (online): Association for Computational Linguistics, Nov. 2021.

[47] M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, and O. Levy, "SpanBERT: Improving Pre-training by Representing and Predicting Spans," *Transactions of the Association for Computational Linguistics*, vol. 8, 01 2020.

[48] M. Paragini and M. Kestemont, "The roots of doubt: fine-tuning a bert model to explore a stylistic phenomenon," in *Proceedings of the Computational Humanities Research Conference 2022 (CHR 2022), 12-14 December, 2022, Antwerp, Belgium*, 2022, pp. 72–91.

[49] X. Zhang, F. Wei, and M. Zhou, "Hibert: Document level pre-training of hierarchical bidirectional transformers for document summarization," *arXiv preprint arXiv:1905.06566*, 2019.

[50] Q. Li, H. Peng, J. Li, C. Xia, R. Yang, L. Sun, P. S. Yu, and L. He, "A survey on text classification: From shallow to deep learning," 2021.

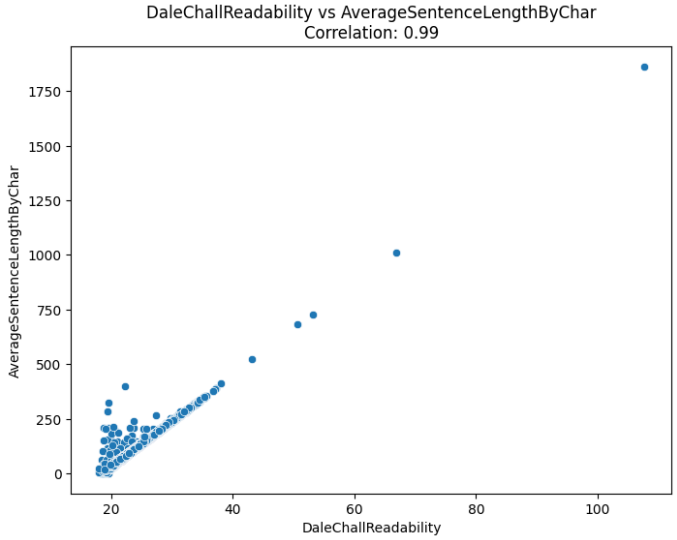
[51] A. Elmahdy, H. A. Inan, and R. Sim, "Privacy leakage in text classification: A data extraction approach," *arXiv preprint arXiv:2206.04591*, 2022.

A Appendix

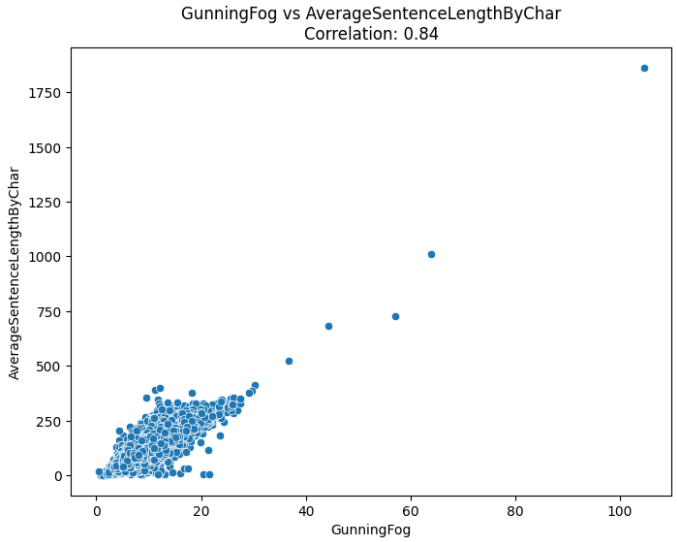
The Appendix Section contains the remaining results related to the study.

A.1 RQ1

This section presents the remaining results related to **RQ1** in Section 4.2. The following Figures show the pairwise *Pearson* correlation between the various features of the textual properties of the text segments.

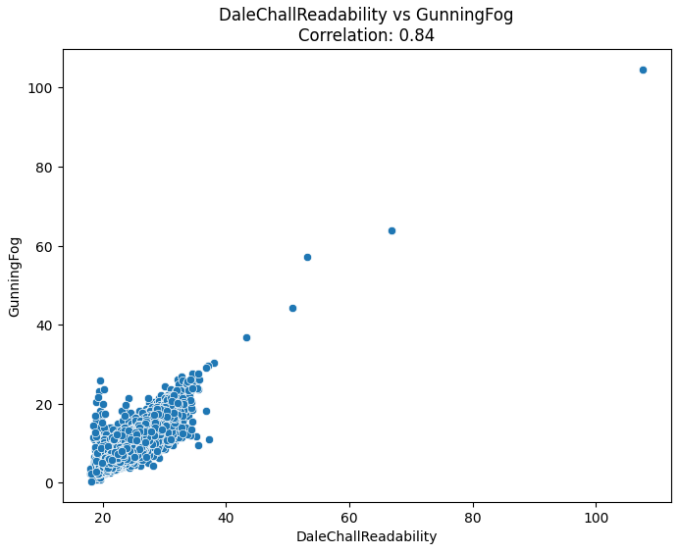


(a) Dale-Chall Readability Vs. Average Sentence Length By Char.

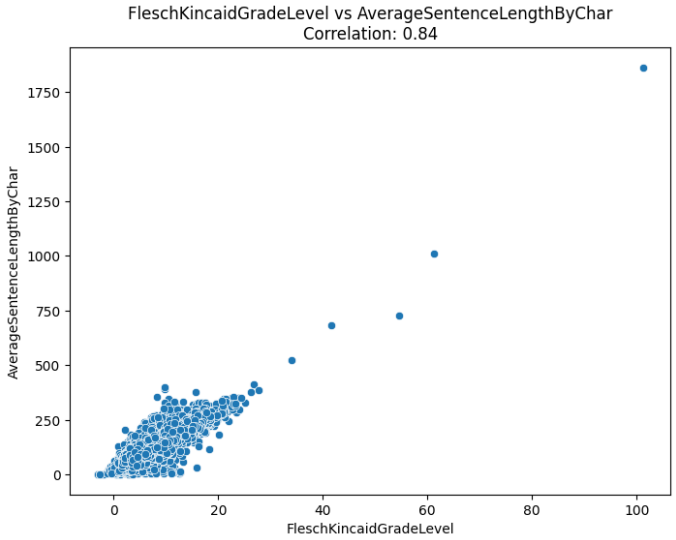


(b) Gunning Fog Vs. Average Sentence Length By Char.

Figure 1.1: Pairwise-Correlated Features (A).

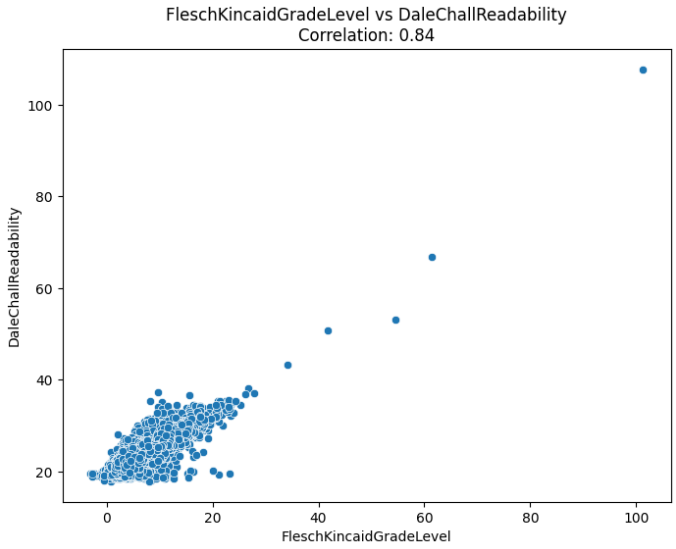


(a) Dale-Chall Readability Vs. Gunning Fog.

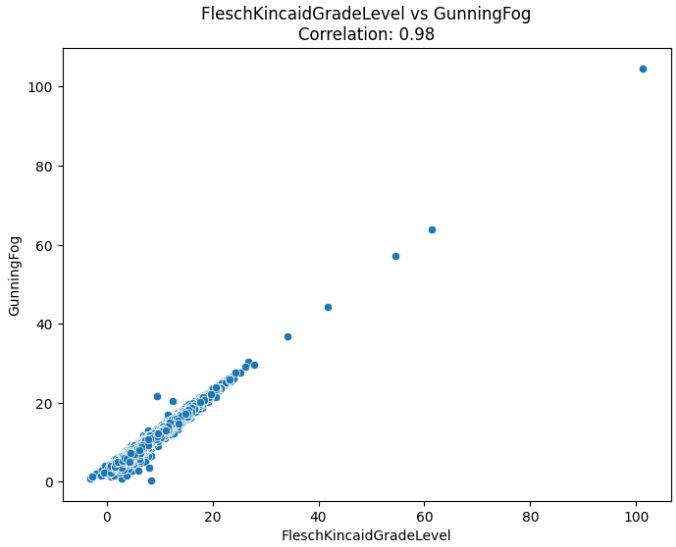


(b) Flesch-Kincaid Grade Level Vs. Average Sentence Length By Char.

Figure 1.2: Pairwise-Correlated Features (B).

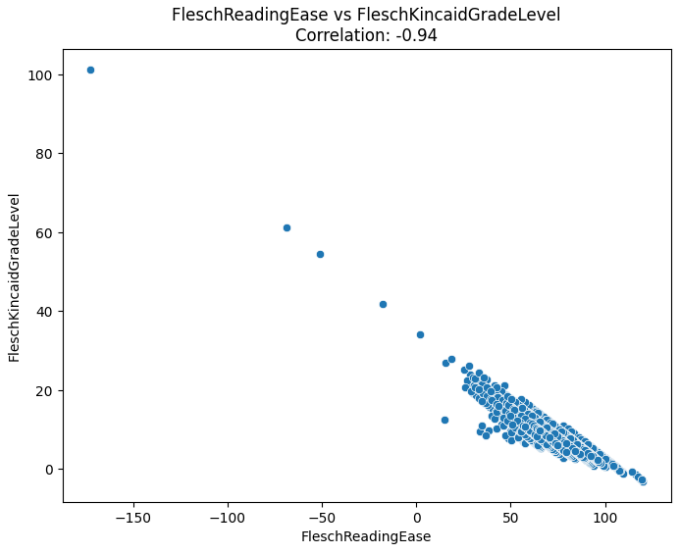


(a) Flesch-Kincaid Grade Level Vs. Dale-Chall Readability.

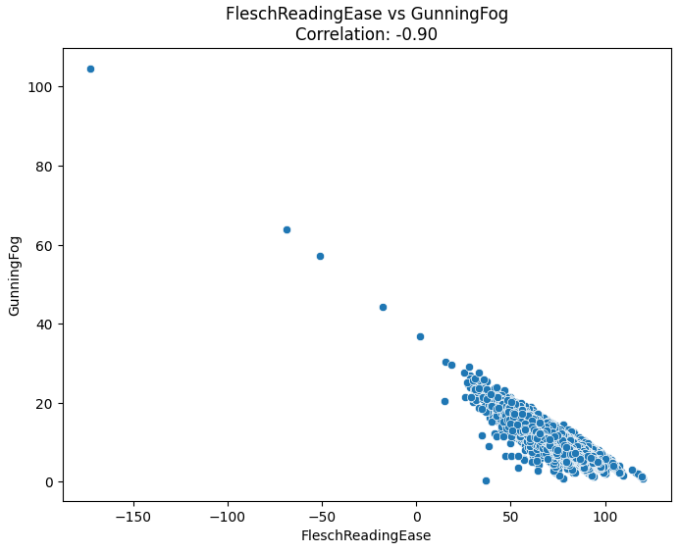


(b) Flesch-Kincaid Grade Level Vs. Gunning Fog.

Figure 1.3: Pairwise-Correlated Features (C).

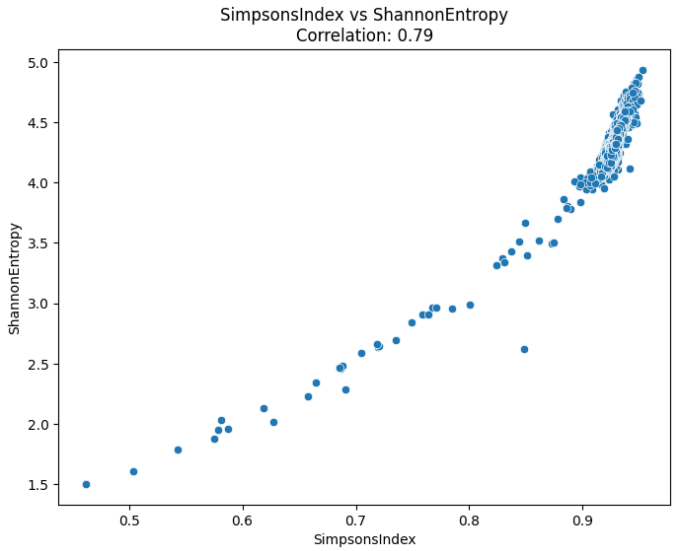


(a) Flesch Reading Ease Vs. Flesch-Kincaid Grade Level.

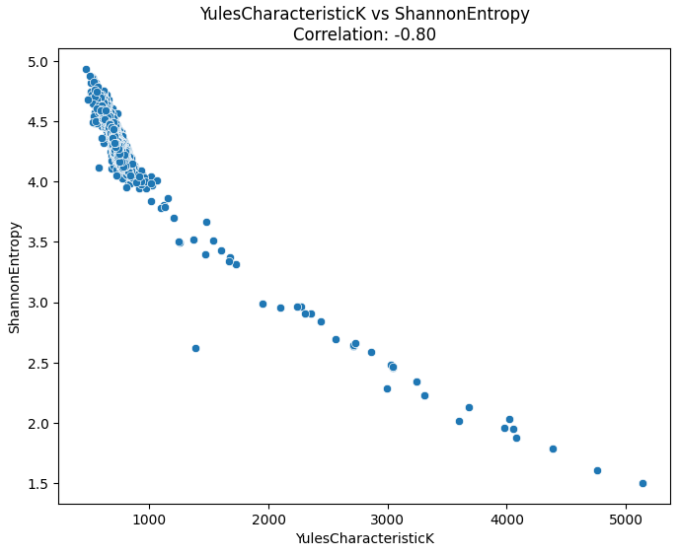


(b) Flesch Reading Ease Vs. Gunning Fog.

Figure 1.4: Pairwise-Correlated Features (D).

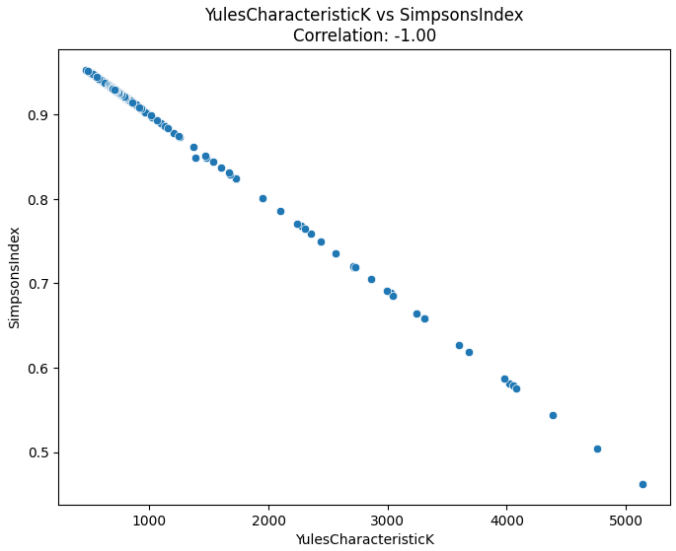


(a) Simpson's Index Vs. Shannon Entropy.

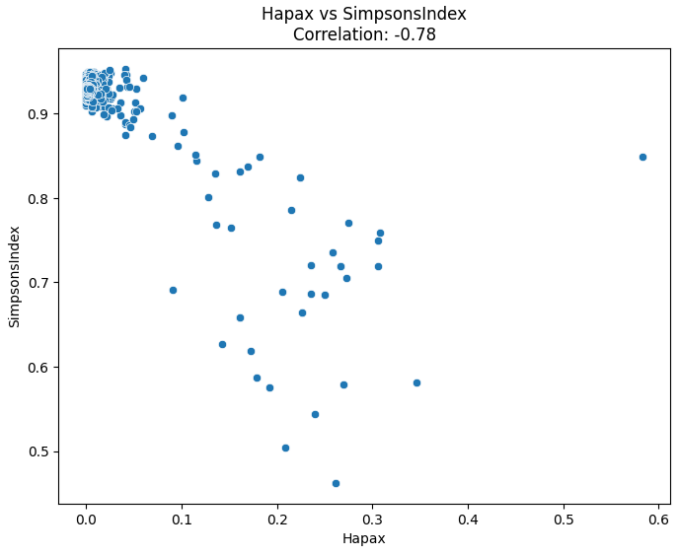


(b) Yule's Characteristic K Vs. Shannon Entropy.

Figure 1.5: Pairwise-Correlated Features (E).

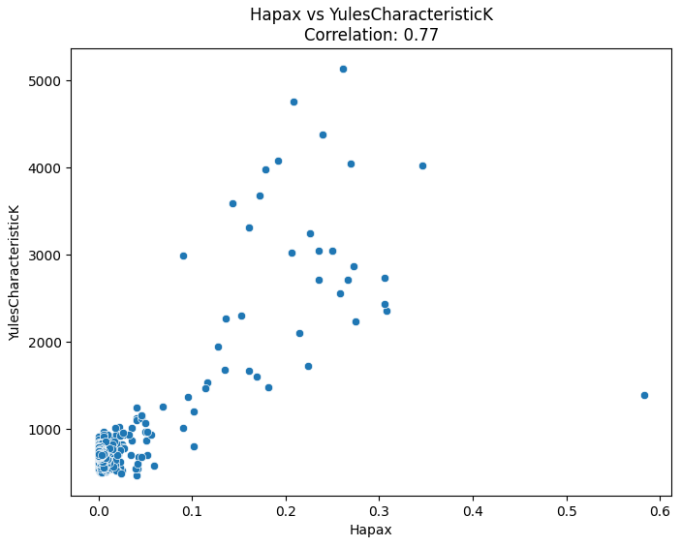


(a) Yule's Characteristic K Vs. Simpson's Index.

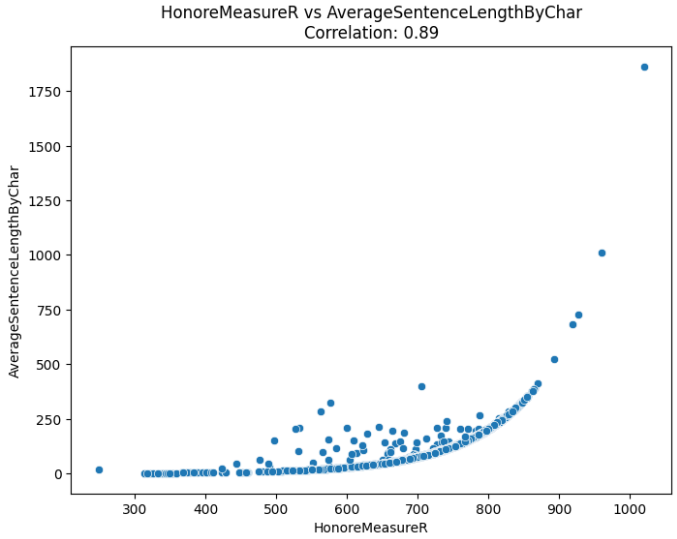


(b) Hapax Legomena Vs. Simpson's Index.

Figure 1.6: Pairwise-Correlated Features (F).

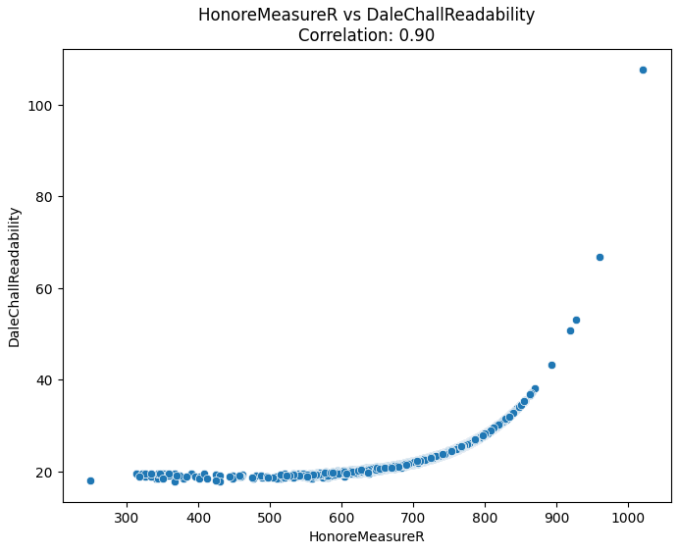


(a) Hapax Legomena Vs. Yule's Characteristic K.

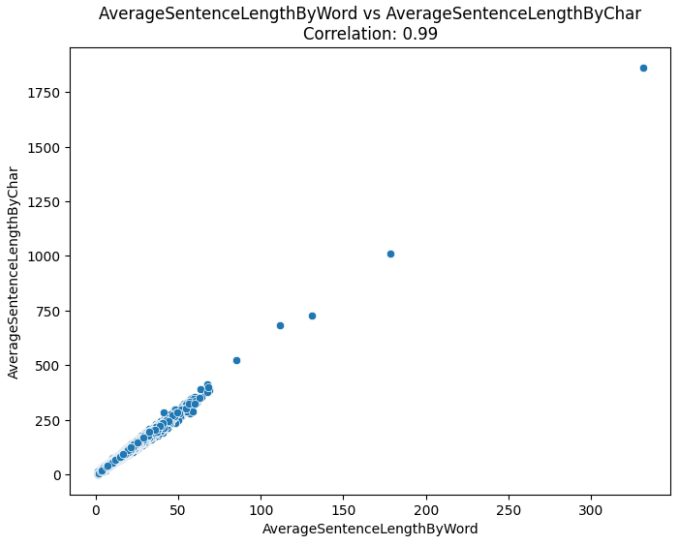


(b) Honore Measure R Vs. Average Sentence Length By Char.

Figure 1.7: Pairwise-Correlated Features (G).

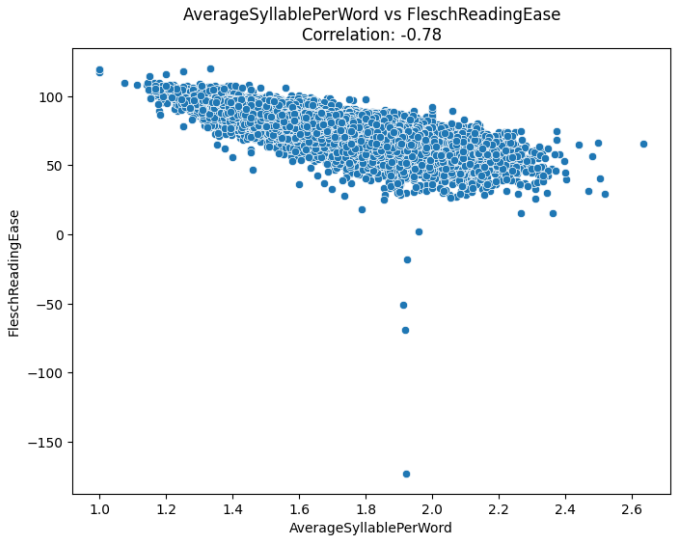


(a) Honore Measure R Vs. Dale-Chall Readability.

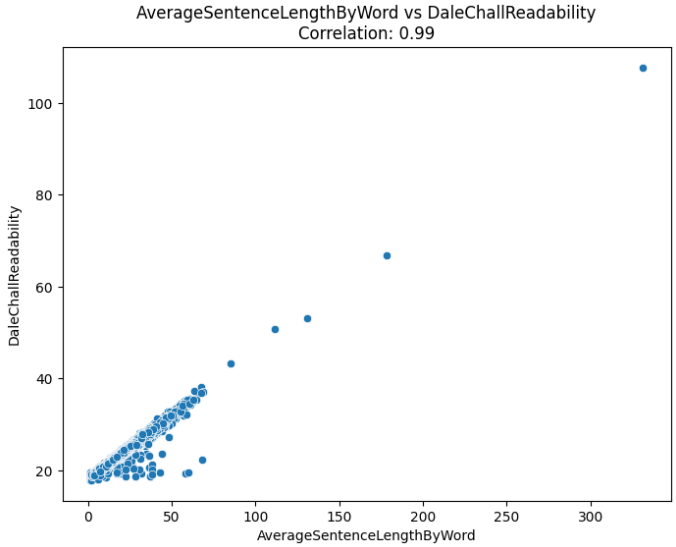


(b) Average Sentence Length By Word Vs. Average Sentence Length By Char.

Figure 1.8: Pairwise-Correlated Features (H).

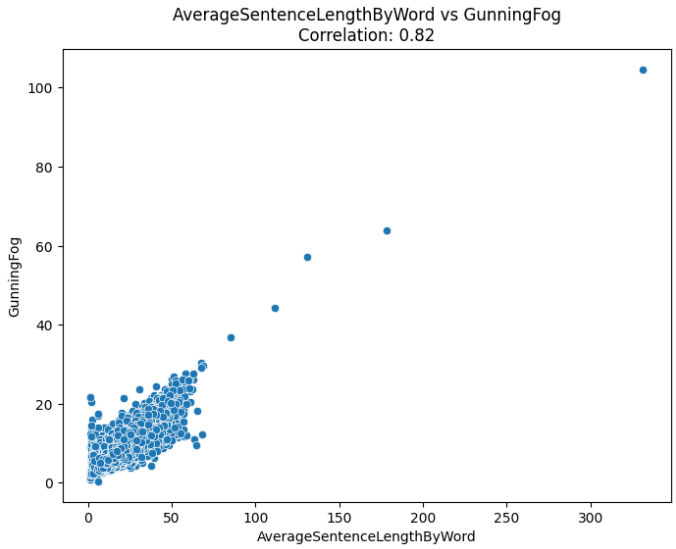


(a) Average Syllable Per Word Vs. Flesch Reading Ease.

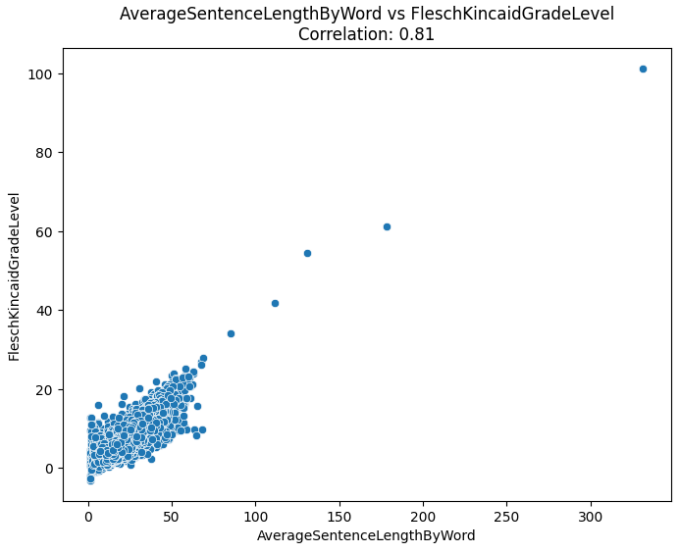


(b) Average Sentence Length By Word Vs. Dale-Chall Readability.

Figure 1.9: Pairwise-Correlated Features (I).

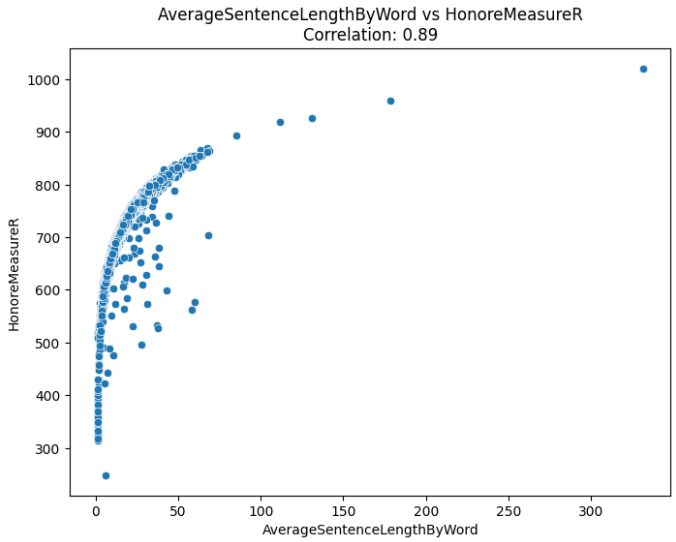


(a) Average Sentence Length By Word Vs. Gunning Fog.

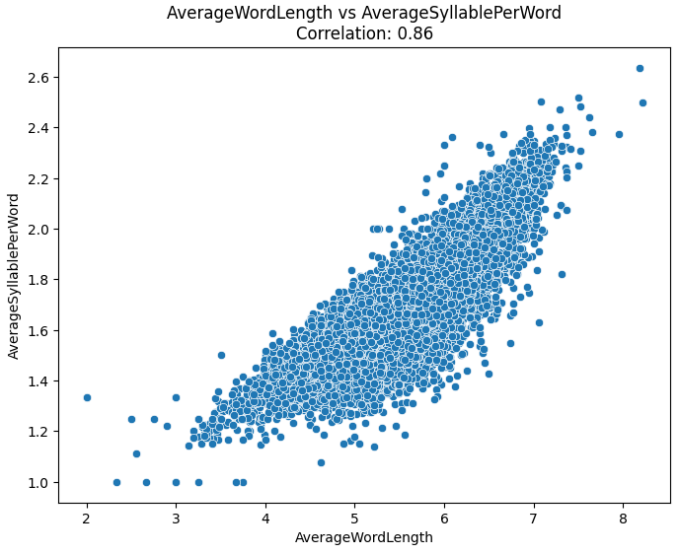


(b) Average Sentence Length By Word Vs. Flesch-Kincaid Grade Level.

Figure 1.10: Pairwise-Correlated Features (J).



(a) Average Sentence Length By Word Vs. Honore Measure R.



(b) Average Word Length Vs. Average Syllable Per Word.

Figure 1.11: Pairwise-Correlated Features (K).

Figure 1.12 visualizes the formation of the clusters' groping the book segments based only on the highly correlated textual features.

Figure 1.12: Clusters Visualization Based On The Highly Correlated Textual Features.

Figure 1.1 represents the orders of the books according to the highly correlated features by computing the average score for each book based on the mean values of the highly correlated features.

Table 1.1: Ranking of Books By Average Score Of Highly Correlated Features.

Book	**Average Correlated Feature Score**
Lady Windermeres Fan	110.6
The Importance Of Being Earnest	111.9
The History Of A Crime	128.5
A Room With A View	128.5
The Picture Of Dorian Gray	129.2
Howards End	129.5
Ninety- Three	129.6
Sentimental Education	130.4
Ozma Of Oz	130.7
Lord Jim	130.9
A Passage To India	131.1
Nostromo	131.3
White Fang	131.4
Heart Of Darkness	131.6
The Sea Wolf	132.3
The Adventures Of Tom Sawyer	132.6
The Lair Of The White Worm	133.3
Bartleby	133.4
What Men Live By	133.8
The Golden Age	134.03
Les Miserables	134.3
War And Peace	134.6

Continued on next page

Table 1.1 –*Continued from previous page*

Book	Average Correlated Feature Score
The Call Of The Wild	134.6
Anna Karenina	134.7
Madame Bovary	135.3
Salammbo	135.8
The Marvelous Land Of Oz	135.9
Dream Days	136.1
The Wind In The Willows	136.5
Moby- Dick	136.7
The Jewel Of Seven Stars	136.9
Oliver Twist	136.99
A Tale Of Two Cities	138.2
Frankenstein	138.4
Dracula	138.4
The Piazza Tales	138.4
Emma	139.2
Middlemarch	139.2
Pride And Prejudice	139.3
Great Expectations	139.5
The Adventures Of Huckleberry Finn	139.9
The Wonderful Wizard Of Oz	140
The Last Man	140.7
The Mill On The Floss	142.1
Sense And Sensibility	142.4
A Connecticut Yankee In King Arthurs Court	143.5
Silas Marner	144.8
Valperga	145.8

A.2 Side Contribution: Exploratory Analysis

This section presents the outcome of the exploratory analysis.

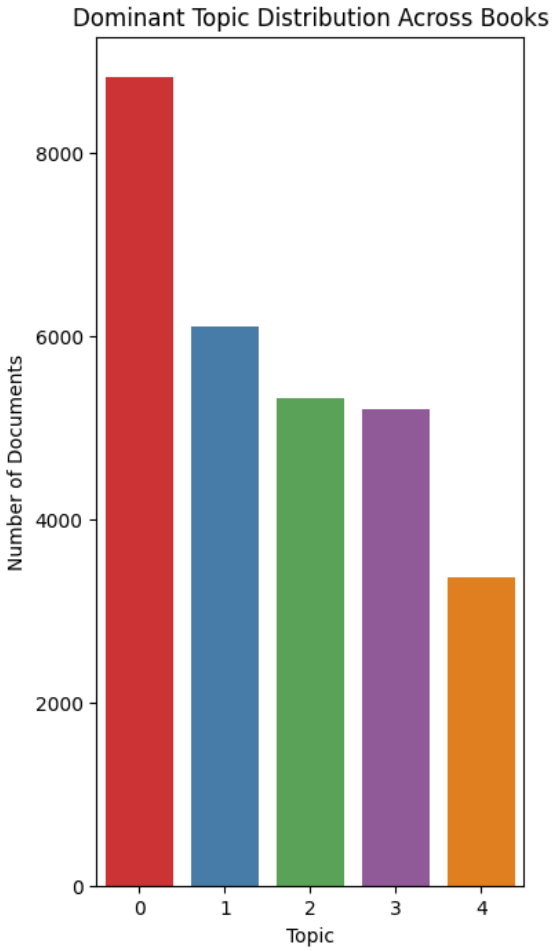


Figure 1.13: Dominant Topic Distribution Across Books.

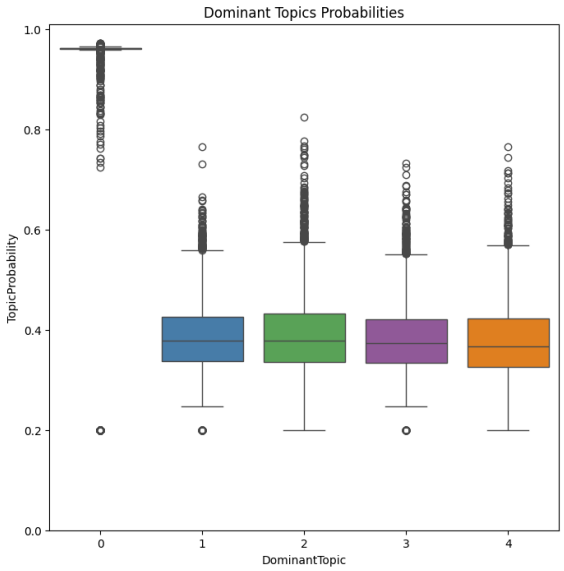


Figure 1.14: Dominant Topics Probabilities.

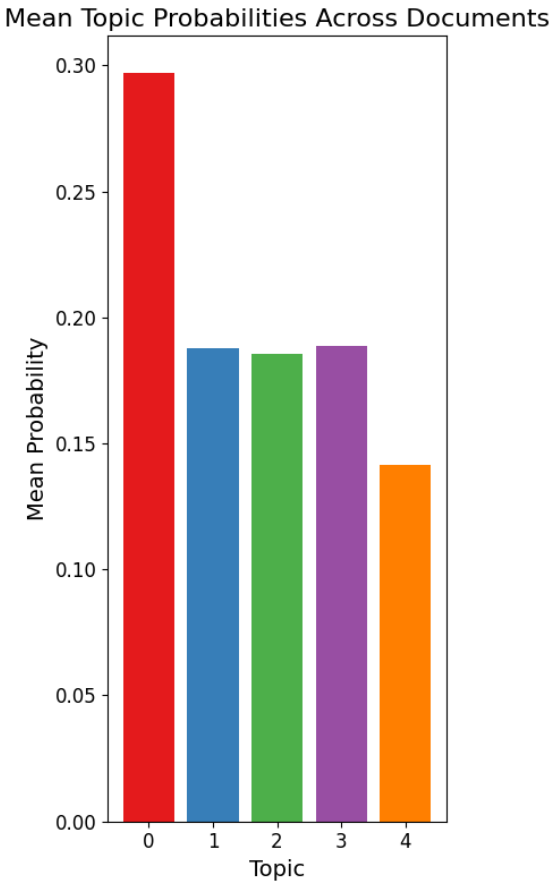


Figure 1.15: Mean Topic Probabilities Across Documents.

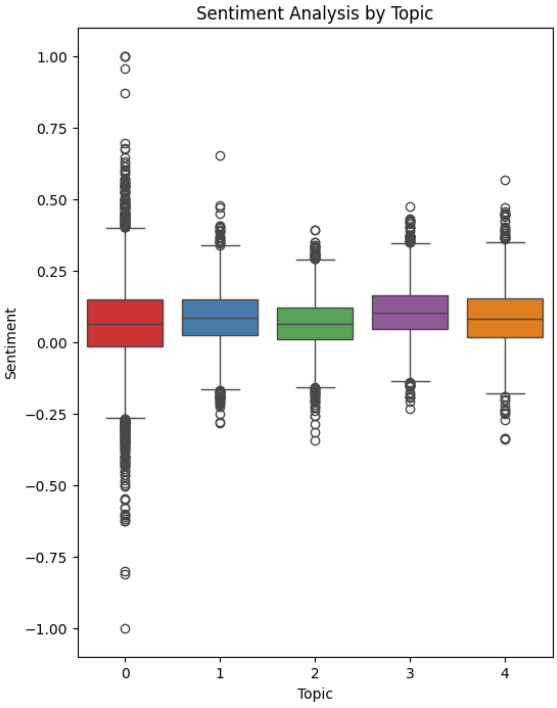


Figure 1.16: Sentiment Analysis By Topic.

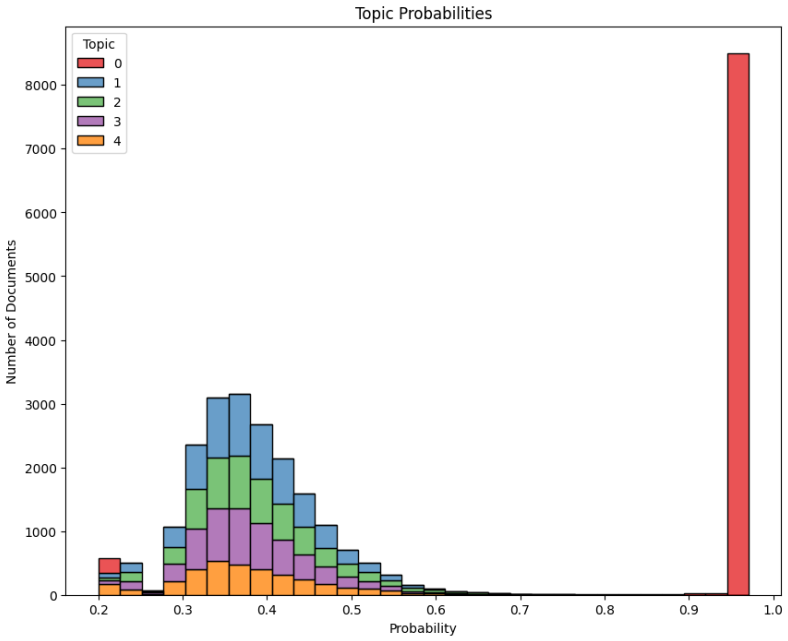


Figure 1.17: Topic Probabilities.

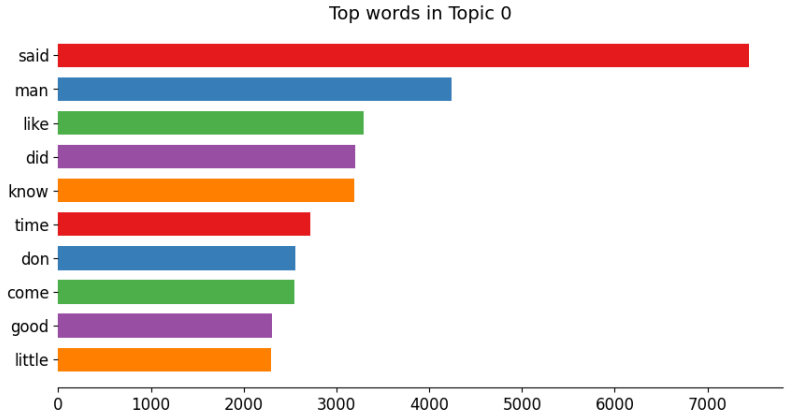


Figure 1.18: Topic words In Topic 0.

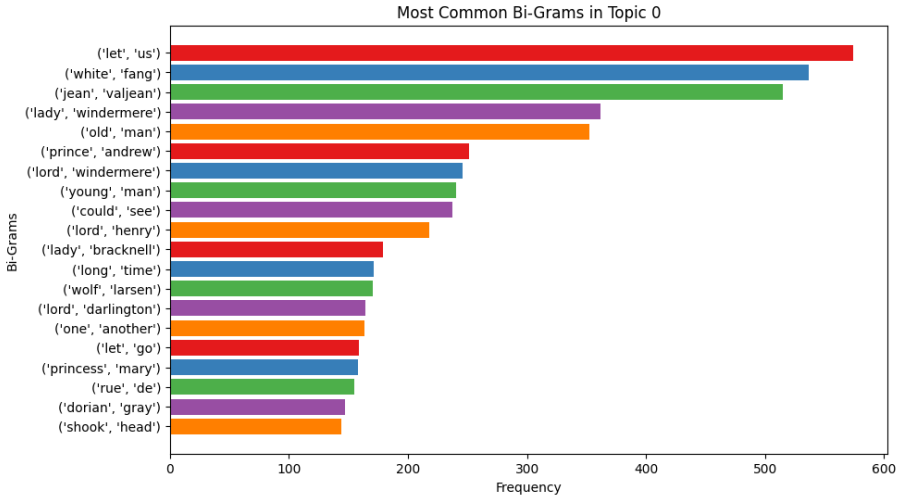


Figure 1.19: Most Common Bi-Grams In Topic 0.

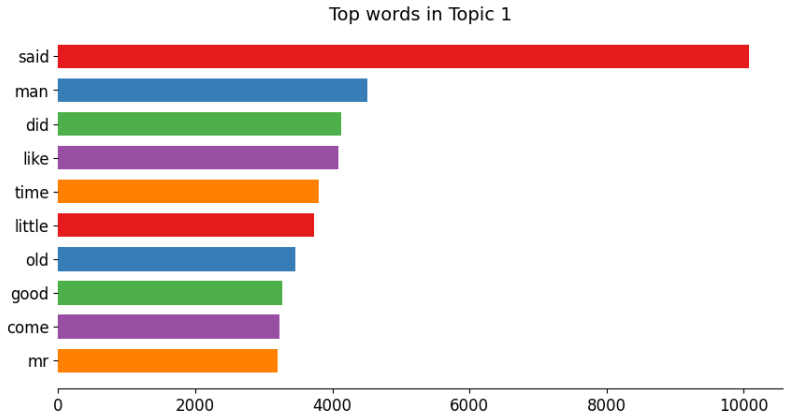


Figure 1.20: Topic words In Topic 1.

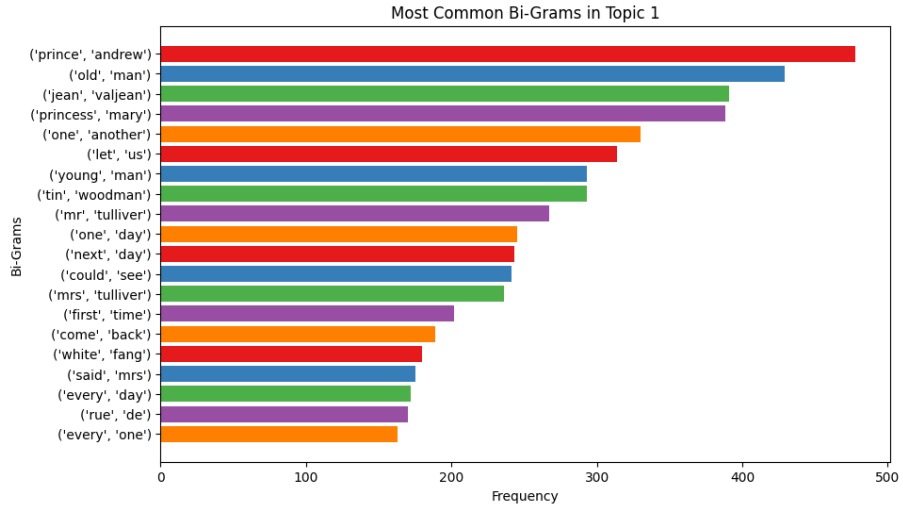


Figure 1.21: Most Common Bi-Grams In Topic 1.

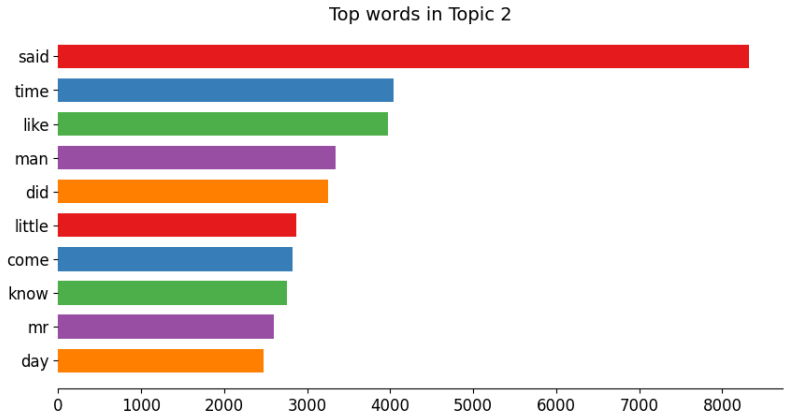


Figure 1.22: Topic words In Topic 2.

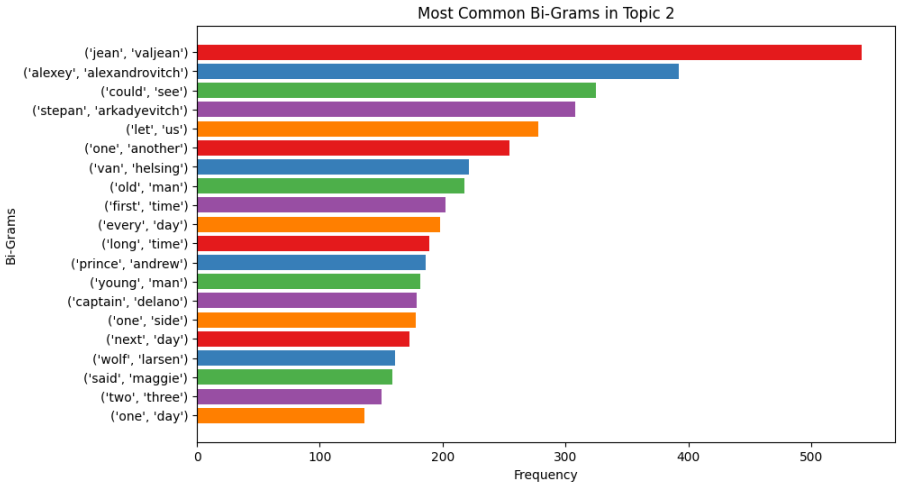


Figure 1.23: Most Common Bi-Grams In Topic 2.

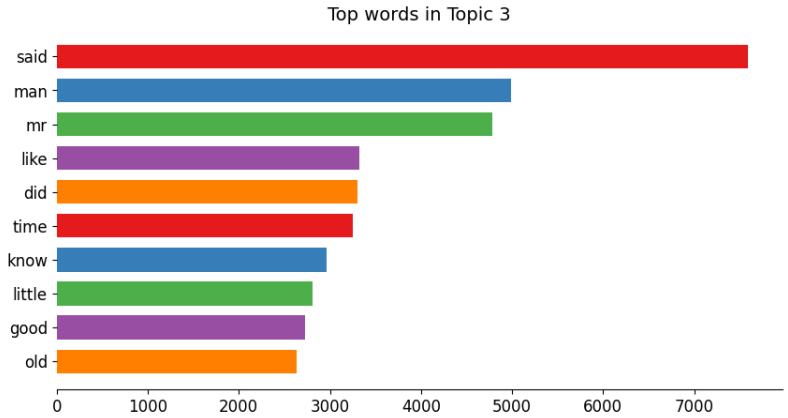


Figure 1.24: Topic words In Topic 3.

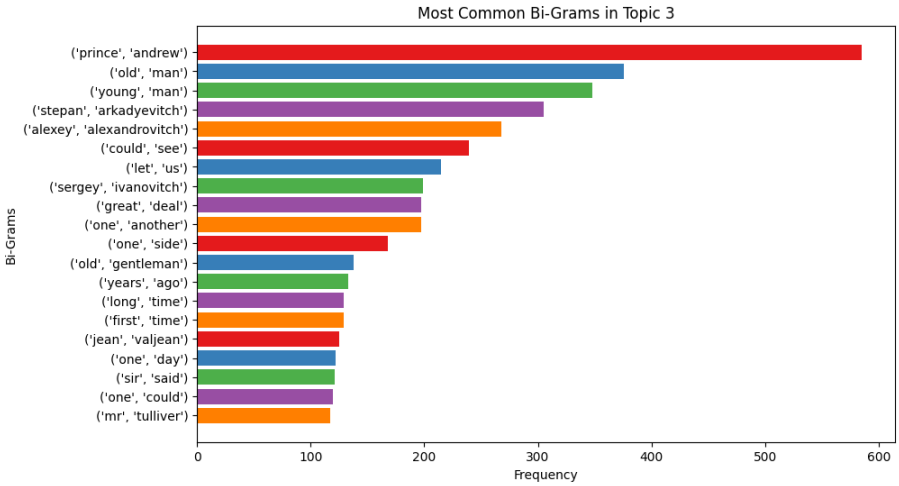


Figure 1.25: Most Common Bi-Grams In Topic 3.

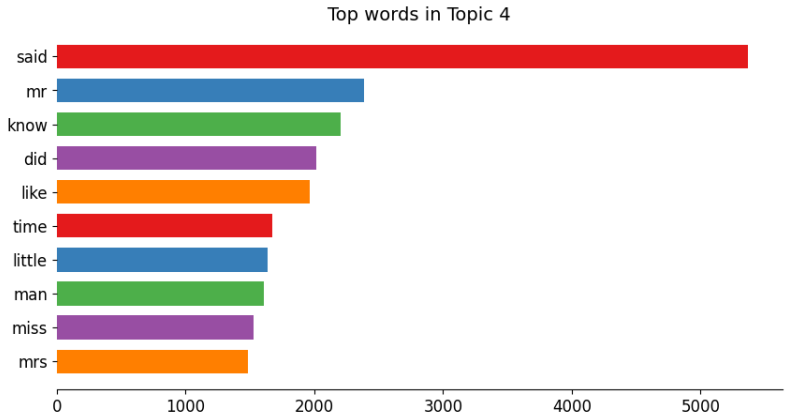


Figure 1.26: Topic words In Topic 4.

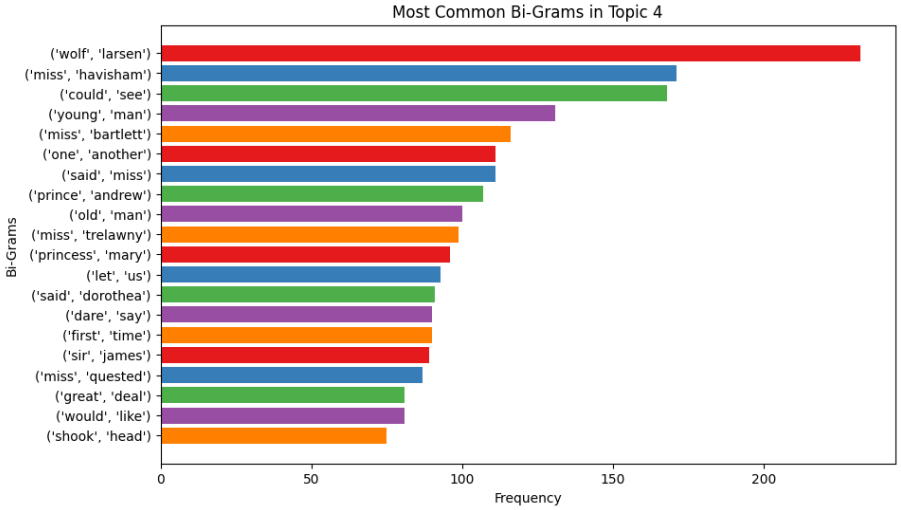


Figure 1.27: Most Common Bi-Grams In Topic 4.

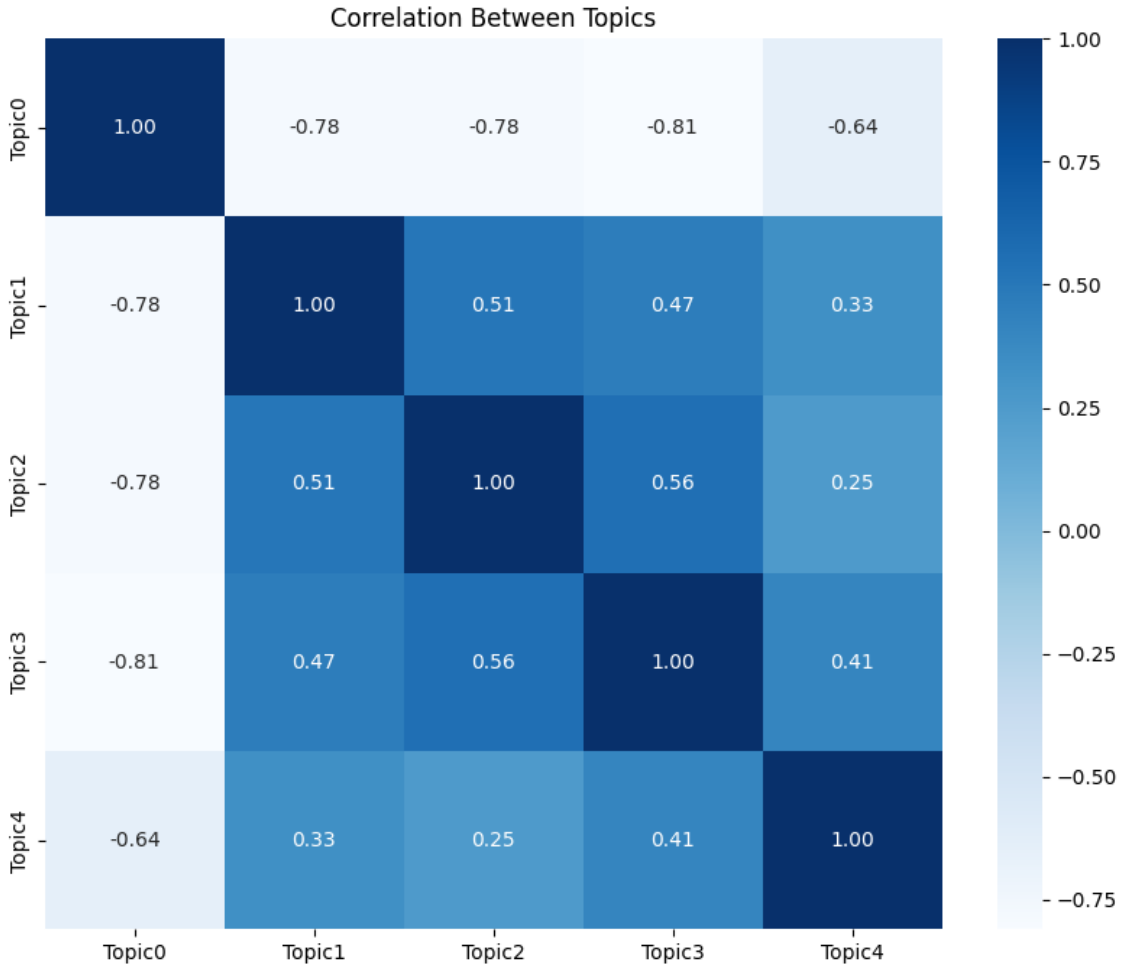


Figure 1.28: Topics Correlations: Pearson's Correlation Coefficient.