Academic Report: Constructing an Open-Set Author Stylometry Dataset

Project: Author Stylometry
Section: Data Preprocessing and Methodology
Your Name

1 Introduction and Objectives

In the fields of computational linguistics and author attribution, a persistent key challenge is the differentiation between "topic" (thematic content) and "style" (stylistic features). Traditional models often erroneously identify thematic keywords (e.g., "detective," "carriage") as discriminators of authorial style. This not only leads to poor model generalization on unseen topics but also renders them incapable of performing the more realistic *Open-Set Recognition* task.

The **primary objective** of this project's data preprocessing phase is to construct a **Topic-Leakage-Proof** and **Open-Set-Ready** anonymized dataset. This forces the model to decouple from content and instead learn deeper stylistic fingerprints, such as syntactic structures, punctuation usage, and function word selection.

2 Dataset Construction Methodology

To achieve the aforementioned objective, I designed and executed a five-stage data processing pipeline:

2.1 Corpus Selection and Open-Set Strategy

- **Data Source:** Project Gutenberg was selected for its extensive repository of public-domain, plain-text corpora.
- **Author Selection:** A total of **12** authors (based on 79 valid works) were chosen and partitioned into two mutually exclusive groups:
  - **9 "In-Set" (IS) Authors:** This group includes `JaneAusten`, `CharlesDickens`, `EdgarAllanPoe`, `MarkTwain`, `ArthurConanDoyle`, `OscarWilde`, `HermanMelville`, `Chesterton`, and `VirginiaWoolf`. Their corpora are used for model **training, validation, and closed-set testing**.
  - **3 "Out-of-Set" (OOS) Authors:** This group includes `MaryShelley`, `NathanielHawthorne`, and `Clara`. Their data is **used exclusively for open-set testing**, and the model is never exposed to them during training or validation.
- **Data Ingestion:** A `Python (requests)` script was developed to automate the ingestion and download of all **85** target works (of which 79 were valid and incorporated).

2.2 Text Cleaning and Anonymization

This stage aims to maximally strip thematic cues, compelling the model to learn more abstract, structural features.

1. **Noise Removal:** Regular Expressions (Regex) were used to batch-process all texts, removing non-authorial noise such as Project Gutenberg headers, footers, copyright notices, tables of contents, and editorial notes.
2. **Stylistic Anonymization:** To mitigate topic bias, we utilized `spaCy (en_core_web_md)` to perform **Named Entity Recognition (NER)** on the entire corpus. All identified proper nouns (Persons, Places, Organizations) as well as dates and numbers were replaced with uniform, anonymized placeholders (e.g., `<PER>`, `<LOC>`, `<NUM>`).

2.3 Segmentation

- **Rationale:** While Transformer models require fixed-length inputs, stylistic features (like sentence length distribution) are dependent on complete sentence structures.
- **Implementation:** `NLTK (punkt)` was used to segment the texts into lists of sentences. These sentences were then **recombined** into text segments ("chunks") with a length between **128 and 512 tokens**. This method satisfies the model's input requirements while maximally preserving local syntactic integrity.

2.4 Leakage-Proof Stratified Splitting

To rigorously prevent data leakage—whereby a model "memorizes" content rather than learning style by seeing parts of the same book in both train and test sets—we implemented a two-stage, *work-based* stratified split:

1. **In-Set Authors (9 authors, 79 works):**
   - **Stage 1:** The 79 works were stratified by author and split 70%/30% into a 'Train' set (55 works) and a 'Test/Val' pool (24 works).
   - **Stage 2:** The 'Test/Val' pool (24 works) was again stratified by author and split 50%/50% into a 'Val' set (12 works) and a 'Test' set (12 works).
2. **Out-of-Set Authors (3 authors, 6 works):**
   - All 6 works were **100% allocated** to the 'Test' set.
3. **Result:** The 'train' and 'val' sets contain only the 9 known authors, while the 'test' set contains all 12 authors, perfectly simulating the open-set test scenario.

2.5 Data Balancing

- **Problem:** The raw, segmented data exhibited extreme class imbalance (e.g., `CharlesDickens` yielded 4,492 samples, while others yielded only a few hundred).
- **Solution:** A **Downsampling** strategy was employed. We established a hard **ceiling of 1,000 samples** per author.
  - For authors with $>$ 1,000 samples (e.g., Dickens, Poe), a random subset of 1,000 samples was retained.
  - For authors with $<$ 1,000 samples (e.g., Twain, Wilde), all available samples were preserved.

3 Final Dataset Characteristics

The final delivered dataset, `author_style_dataset_OPENSET.csv`, contains **9,781** anonymized text segments.

3.1 Distribution by Split

Table 1: Distribution of Dataset Splits

Metric	**train**	**val**	**test**
Sample Count	5,595	1,409	2,777
Proportion	57.2%	14.4%	28.4%
Author Count	**9** (In-Set Only)	**9** (In-Set Only)	**12** (All Authors)

3.2 Distribution by Author

To mitigate severe class imbalance observed in the raw segmented data (e.g., `CharlesDickens` at 4,492 samples), a downsampling strategy with a hard ceiling of **1,000 samples per author** was implemented.

For the six over-represented authors (`Dickens`, `Doyle`, `Austen`, `Woolf`, `Poe`, `Melville`), a random subset of 1,000 samples was retained. For the remaining authors, all available samples were preserved. This resulted in a significantly more balanced dataset: `OscarWilde` (858 samples), `Chesterton` (806), and `MarkTwain` (792).

Critically, this balancing was also applied to the "out-of-set" authors, whose data is present only in the test set. Their full sample counts were retained: `NathanielHawthorne` (723), `Clara` (386), and `MaryShelley` (216). This process ensures that no single author disproportionately influences the model, while preserving the entirety of the scarcer data.