Authorship Attribution using Poisson Probabilities: A Case Study on Jane Austen

Utkarsh Bharadwaj
Indian Statistical Institute, Delhi

April 3, 2026

Abstract

This report outlines an initial exploration into identifying the authorship of historical texts using basic statistical modeling. By analyzing the frequency of subconscious "filler" words, we attempt to model writing styles using a Poisson distribution. Specifically, we compare texts by Jane Austen against a combined corpus of other prominent 19th-century authors. This document serves as a baseline methodology to facilitate further discussion and refinement of the underlying code and statistical assumptions.

1 Introduction

The primary motive of this project is to explore whether we can reliably identify a specific author's writing style using basic statistical methods. Because authors use common "filler" or "function" words (like *in, that, it, is*) subconsciously, the frequency of these words acts as a unique fingerprint. By modeling the occurrence of these words using a Poisson distribution, we aim to calculate the probability that an unknown text was written by Jane Austen compared to a pool of other 19th-century english authors.

2 Dataset

The data for this analysis consists of plain text files (.txt) of classic literature, divided into two main categories:

- **Target Author (Jane Austen):** *Emma, Northanger Abbey, Pride and Prejudice,* and *Sense and Sensibility.*
- **Comparison Group (Not Austen):** A selection of works by other prominent authors of similar eras, including Charles Dickens (*A Tale of Two Cities, Bleak House*), Mary Shelley (*Frankenstein*), Charlotte Brontë (*Jane Eyre*), Emily Brontë (*Wuthering Heights*), George Eliot (*Middlemarch, Silas Marner*), and Sir Walter Scott (*Ivanhoe, Waverley*).

3 Implementation and Methodology

The initial analysis was implemented in R, following these core steps:

1. **Text Processing:** The texts were read into R and cleaned using regular expressions. All text was converted to lowercase. Apostrophes were removed entirely (to prevent splitting words like "it's"), and all non-alphabetic characters were replaced with spaces. This standardized the data into simple vectors of words.

2. **Model Training:** For both the "Austen" and "Not Austen" datasets, all books within the respective category were merged into one massive vector of words. We then calculated the rate of occurrence for 15 specific filler words per 1,000 words of text. During this step, we added a small baseline value to every word's count. This is a mathematical failsafe to ensure no word has a probability of exactly zero, which would cause the Poisson distribution calculations to fail.

3. **Evaluation:** To test an unknown text, the script cleans the text using the exact same rules and counts the occurrences of our chosen filler words. Using the '`dpois()`' function in R, it calculates the log-likelihood that the observed counts came from the Austen model versus the Comparison model.

4. **Result:** By summing the log-likelihoods of all filler words, we avoid numerical underflow and generate a final percentage representing the probability that Austen authored the text.

4 Limitations and Open Questions

As this is an exploratory project, the current model has several limitations that serve as starting points for improvement:

- **Data Aggregation:** We combined all the textbooks for Austen into one single file (and did the same for the "Not Austen" group). This averages out the writing style and completely ignores the stylistic variations that exist between individual books.

- **The Independence Assumption:** The model assumes that the occurrence of one filler word is strictly independent of another. In natural language, this is false (e.g., the word "it" might frequently be followed by "is"). This can lead to the model being overconfident in its final probabilities.

- **Aggressive Text Cleaning:** By stripping out all numbers and punctuation, we potentially lose valuable stylistic identifiers. For instance, how an author uses em-dashes or semicolons could be a strong indicator of authorship that our current model ignores.