

# Author Recognition

KARTHIK KADEWADI PES1UG21CS271

KALLESCH ACHAR KG PES1UG21CS266

# INTRODUCTION TO PROBLEM DOMAIN

- ▶ Author recognition, also known as authorship attribution or author profiling, is the process of identifying the authors of given texts based on their writing style and linguistic patterns. This task has various applications, including plagiarism detection, forensic linguistics, and social media analysis. By analyzing the unique characteristics of an author's writing style, it becomes possible to attribute authorship to previously anonymous or disputed texts.


# Problem Statement

- ▶ The objective of this project is to develop a machine learning model that can accurately identify the authors of given texts based on their writing style and linguistic patterns. Given a dataset of texts written by different authors, the model should be able to predict the most likely author of a given text.

# Brief Literature Review:

- ▶ Several studies have been conducted in the field of authorship attribution and stylometry. Most approaches rely on extracting various linguistic and stylistic features from the text, such as word frequency, sentence length, and syntactic structures. Traditional machine learning algorithms, such as support vector machines and random forests, have been commonly used for author recognition tasks. Recent advancements in deep learning have also shown promising results in this domain, particularly with the use of recurrent neural networks and convolutional neural networks.

# Dataset: The Gutenberg Project



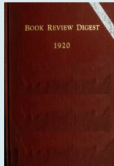
[About](#) [Search and Browse](#) [Help](#)

[Go!](#) [Donation](#) [PayPal](#)


## Welcome to Project Gutenberg

**Project Gutenberg is a library of over 70,000 free eBooks**


Choose among free epub and Kindle eBooks, download them or read them online. You will find the world's great literature here, with focus on older works for which U.S. copyright has expired. Thousands of volunteers digitized and diligently proofread the eBooks, for you to enjoy.




**The Cumulative Book Review**  
1920




**Rämekorven viinakuninkaat**  
Rusko korvasta  
Veikko Korpunen



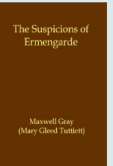
**MIREILLE DES TROIS RAISINS**  
PIERRE LA MAZIERE  
F. G. FORTIN




**RUBE BURROW, KING OF OUTLAWS**  
BY RUBE BURROW  
D. W. FARR




**THE MEDIAEVAL STAGE**  
VOLUME 1  
E. V. Rieu



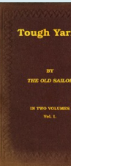
**The Suspicions of Ermengarde**  
Maxwell Gray  
(Mary Cloud Taylor)



**WORTH HIS WHILE**  
AMY E. BLANCHARD



**TRANSPLANTED**  
BY GERTRUDE ATHERTON  
FRANKLIN HORN



**Tough Yarn**  
BY THE OLD SAILOR  
OF THE OLD SAILOR

*Some of our latest eBooks* [Click Here for more latest books!](#)

# Literature review

Detection of changes in literary writing style using N-grams as style markers and supervised machine learning | PLOS ONE

Principal Component Analysis (PCA) and Latent Semantic Analysis (LSA)

Factors considered: characters, words, Part-Of-Speech (POS) tags and syntactic relations n-grams

Focused on chronology, an author's style changing over time

Dimensionality reduction to see the most important factors

The dimensionality reduction process can be defined as follows: Given a matrix  $A$  of  $m \times n$ , where  $n$  is large; it is often desirable to project the  $m$  lines to a smaller dimensional space, to a matrix of  $m \times k$ , with  $k < n$ , where  $k$  represents the new dimensions of the matrix. It is difficult to determine the appropriate value of  $k$ , because it depends on the dataset. A common heuristic for estimating  $k$  involves setting a threshold. In this analysis, experiments were carried out using two strategies: (1) selecting  $k$  dimensions where  $k$  is the number of samples in the training set and (2) selecting the  $k$  most informative features (commonly  $k = 2$ ).

# Literature review

[1906.03072.pdf \(arxiv.org\)](#)

Investigating Writing Style Development in High School

Siamese Neural Network

The network relies only on character level inputs. The basic philosophy behind the network is to a) encode the two texts in some space using a replicated encoder network with shared weights, and b) compare the two texts in this space.

The encoder network in the encoding part of our network consists of a character embedding (using ReLU activation functions), followed by two different convolutional layers (CONV): one using kernel size  $k = 8$  and  $n = 700$  filters, and one using  $k = 4$  and  $n = 500$ , each followed by global max pooling layers (GMP). • In the comparison part of the network, the MERGE layer first computes the absolute difference between the outputs of the two encoder networks. Afterwards, four dense layers (DENSE) with 500 neurons each are applied, using ReLU for activation function and with a dropout of 0.3. Finally a two neuron softmax layer is used to normalize the output.

# Proposed Solution

- ▶ Our proposed solution involves collecting a diverse dataset of texts written by different authors and preprocessing the data to extract relevant features. We will experiment with various machine learning models, including logistic regression, support vector machines, and neural networks, to develop an accurate author recognition system. The model will be trained and evaluated using standard evaluation metrics, such as accuracy, precision, recall, and F1-score. Additionally, we will explore techniques for feature selection and model optimization to improve performance.