CMPE 255 - Data Mining
Fall 2021
Group Project

Project Title: Author Identification based on stylometry

By
Group 10
Harshavardhana Reddy Namburu- 015920775
Niousha Noshiravani - 015963493
Pavan Karthik Gollakaram - 015945670
Teja Ganapati Jaddipal - 015957526

Project Description

Authorship identification can be described as attributing text documents to authors based on stylometric analysis. Based on the fact that each author has a unique writing style, we are aiming at identifying the author for certain text documents. Leveraging the vast availability of digitized books, magazines, and articles in recent times, we propose this Author Identification model in order to avoid malicious duplication of either academic or professional material.

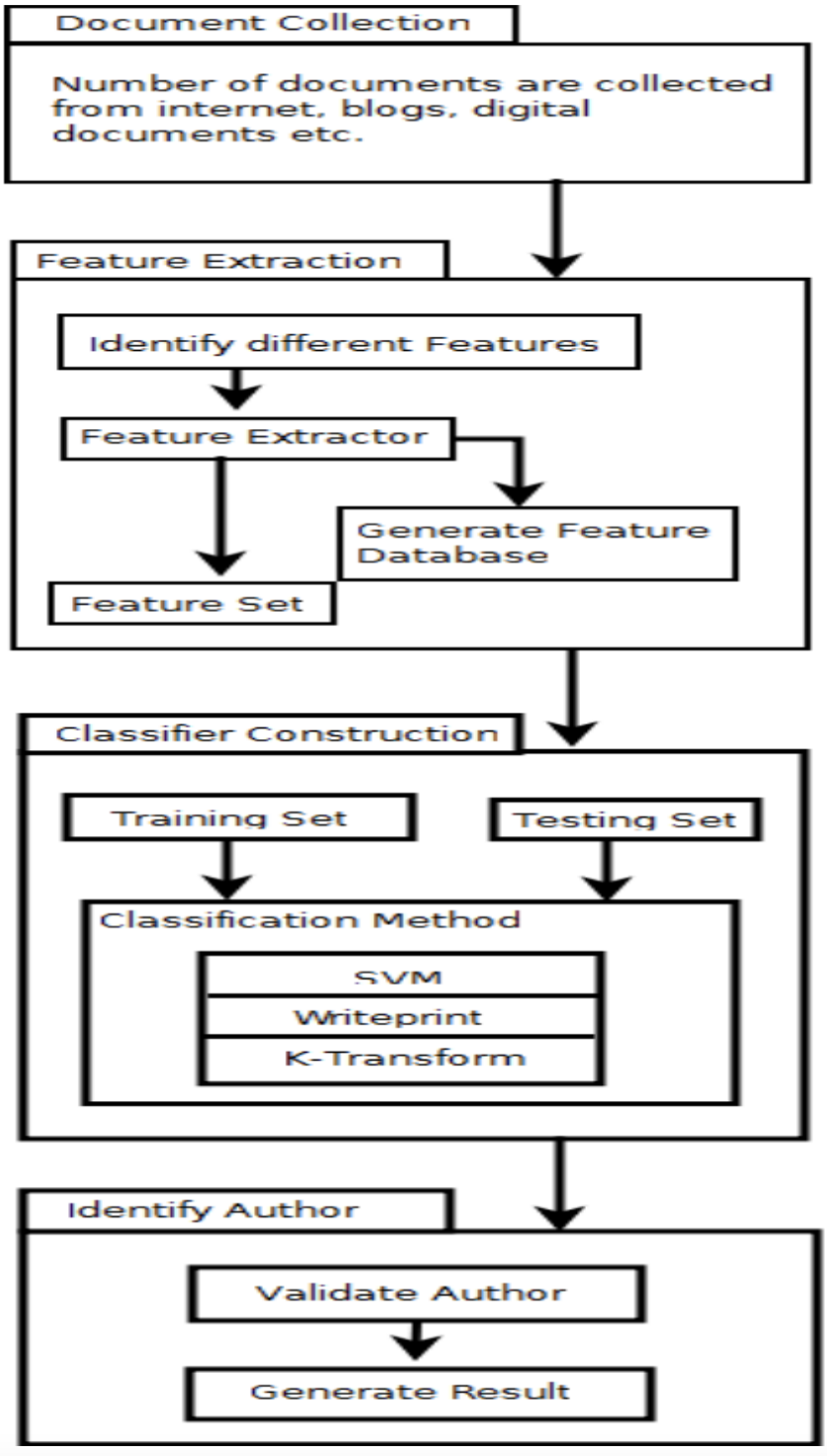
Proposed Methodology

The dataset we opted for is the UCI's Reuter_50_50 dataset with 5000 instances and 10000 attributes. The dataset is divided into train and test sets. There are 50 authors and 100 documents for each author with snippets of their writings. The training set includes 50 documents, and the test set has 50 documents.

We are following a stylometry based approach where various features of the text documents such as stylistic and linguistic features [1] are extracted and compared. The document data will be preprocessed to remove stop words, words with shorter length and lemmatization will be used to retain only root words. Punctuation marks which contribute to uniqueness like En dash, Em dash, colons will be retained and rest like full stops and commas will be removed.

Feature extraction steps include analyzing the writing style of each author and creating a feature set. These features include but are not limited to frequently used words, word patterns, sentence format etc. With numerous classification techniques available at our disposal, we would be using trial and error methods to compare one and other to obtain the most efficient and optimal model. Precision, Recall and F1 scores will be used for model evaluation.

Project flow diagram [1]


Document Collection
Number of documents are collected from internet, blogs, digital documents etc.
Feature Extraction
Identify different Features
Feature Extractor
Generate Feature Database
Feature Set
Classifier Construction
Training Set
Testing Set
Classification Method
SVM
Writeprint
K-Transform
Identify Author
Validate Author
Generate Result

Dataset: https://archive.ics.uci.edu/ml/datasets/Reuter_50_50#

References:

[1]. Authorship Analysis and Identification techniques: A review, International Journal of Computer Applications(0975-8887)
https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.402.9751&rep=rep1&type=pdf

[2]. Chen Qian, Tianchang He, Rao Zhang, Department of Electrical Engineering, Stanford University, Stanford.
https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1174/reports/2760185.pdf