Who wrote this : a framework for French novelist identification
Machine Learning for Natural Language Processing 2020

Romain Avouac
ENSAE
romain.avouac@ensae.fr

Margot Eteve
ENSAE
margot.eteve@ensae.fr

Abstract

Identifying and attributing authorship of a document can be beneficial for multiple applications, such as plagiarism detection and bibliometrics. In this project, we develop a NLP pipeline which performs authorship identification using short texts from French 19th century main novelists.

1 Problem Framing

Being able to identify authorship of a document is beneficial for a wide range of applications. It can prove very valuable from an historical perspective, since archives worldwide are full of documents whose authorship is not known with certainty. Furthermore, multiple plagiarism cases in literature could be solved with such an algorithm ; for instance, the authorship of some of Moliere or Shakespeare works has been debated since the 19th century[1]. Against that background, this project aims at developing a natural language processing (NLP) pipeline for authorship identification. Specifically, we focus on the task of identifying authorship of short texts written by French novelists based on their literary style.

2 Experiments Protocol

Data. We choose to limit our experiment to a selection of ten French novelists from the 19th century[2]. Several reasons motivate this choice. First, in order to develop an algorithm that actually distinguishes authors based on their writing style, we need to select authors from a similar time period. The 19th century features coherent and identifiable litterary movements, whereas the 20th century litterary is much more scattered for instance. Besides, the language used in 19th century books appears close enough to contemporary French, enabling the use of embeddings pre-trained on modern corpora. Finally, 19th century books are now in the public domain, and importantly directly available in digital format thanks to the *Project Gutenberg*[3]. For each author, three representative books have been selected : two for the training set and one for the test set. This clear separation between train and test sets ensures that we are not using a given book pecularities (e.g character or place names) for prediction. Each book has been cut into paragraphs and each paragraph has been associated its author as label. This unit of analysis has been chosen because it is large enough that it can contain substantial statistical information and yet small enough that it can be processed easily by most NLP algorithms.

Preprocessing. A major advantage of using data from the *Project Gutenberg* is that it is highly normalized and virtually noiseless. It thus requires very little preprocessing to be directly usable in a machine learning pipeline. First, we perform a tokenization step. However, as we seek to distinguish authors based on their writing style, results can be highly sensitive to the choices made at this step. For instance, the amount of punctuation and the way it is used can be very distinctive of an author's style, yet it is generally removed by standard tokenizers. In order to ensure robustness of the results, we replicate our analysis using various tokenization heuristics. We also experiment with stemming and lemmatization techniques to determine whether normalizing data can improve performance.

[1] See here and here for more details.

[2] Zola, Maupassant, Daudet, Stendhal, Balzac, Flaubert, Hugo, Dumas, Vigny and Verne.

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

Models. .

Training.

Evaluation. The most natural metric to evaluate the performance of a classification problem is accuracy, i.e. the rate of rightly classified instances. However, this metric is not suitable in the case of imbalanced classes, which is the case here since the number of paragraphs per book varies significantly. Since we have no reason *a priori* to favor either precision or recall, we use the F1 metric, which formally corresponds to the harmonic mean of both metrics. Thus, a given predictor will produce a good score only if *both* recall and precision are high enough. Finally, since we are dealing with a multiclass classification task, we have to decide an aggregation procedure to get a global metric. Here, we choose the micro-averaged F1 score, which consists in computing the F1 metric globally by counting the total numbers of true positives, false negatives and false positives. We also provide a more qualitative analysis of performance by looking at the confusion matrix.

3 Results

Results are presented in this Colab notebook[4].

4 Discussion

[4] https://nlp-ensae.github.io/