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 literary movements, whereas the 20th century literary 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.

Models. We compare the performance of several models on the classification task. Our base-

[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/

line consists in a linear classifier (logistic regression) trained on the TF-IDF weights matrix. All the other approaches we leverage are based on continuous word representations. The main challenge in this case is the computation of document vectors – paragraph vectors in our case. Since there aren't actual theoretical guidelines to do so, we employ various models and heuristics. First, we use a fully unsupervised technique : paragraph vectors are computed as a simple average of `FastText` [1] pretrained French word vectors. Then we implement `doc2vec` [2], a supervised approach designed to learn word and document vectors jointly. For these two approaches, classification is also performed by training a linear classifier on the paragraph vectors matrix. Finally, we leverage transformer models, which generally achieve state-of-the-art results in most NLP tasks. Specifically, we use the `CamemBERT` model [3], a `BERT` model pretrained on a large French corpus. The model is fine-tuned in a supervised way through a simple dense layer with softmax activation.

Training. All models are trained in a Google Colab environment with a CPU. Both the NLP and the classification models are fine-tuned using grid search on a validation set which consists in half the test set. The other half is used for final evaluation.

Evaluation. We deal with imbalanced classes – the number of paragraphs per book varies significantly – so accuracy is not relevant as an evaluation metric . As there are no *a priori* reason to favor either precision or recall, we use the F1 metric, which puts equal emphasis on both metrics. Since we deal with is a multi-class classification task, we have to decide on an aggregation procedure to get a synthetic quality metric. 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 analyzing the confusion matrix.

3 Results

Results are presented in this Colab notebook[4].

4 Discussion

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

References

[1] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016.

[2] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196, 2014.

[3] Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. Camembert: a tasty french language model. *arXiv preprint arXiv:1911.03894*, 2019.