Objectiver l'attribution d'auteur et la comparaison de documents: apports du machine learning à l'analyse de corpus

Julien Longhi*[1,2,3], Dieudonné Akpo*†[1], Jérémy Demange*‡[1], and Mohamed Boumrar*§[1]

[1]Institut des humanités numériques – Université de Cergy Pontoise – France
[2]Institut Universitaire de France – Ministère de l'Education nationale, de l'Enseignement supérieur et de la Recherche, Ministère de l'Education nationale, de l'Enseignement supérieur et de la Recherche – France
[3]Laboratoire AGORA – Université de Cergy Pontoise : EA7392 – France

Résumé

Selon Etienne Brunet (2002 : 1, à propos des textes), " les problèmes d'attribution ou de datation sont les plus épineux qui soient ". En effet, " même lorsqu'une distance paraît établie solidement entre deux textes, on ne sait pas toujours à quoi la rattacher. À l'auteur ? À l'époque ? Au sujet traité ? " (p.2). La textométrie a longtemps été un bon moyen d'aborder la question de l'attribution d'auteurs, notamment d'un point de vue différentiel, mais des travaux récents (Brunet et al. 2021 : 75) ont montré que " l'approche du deep learning [...] laisse espérer une sensibilité plus grande et plus exacte à la spécificité des textes ". Des différences méthodologiques fortes existent néanmoins entre ces approches (constitution des corpus, établissement des " scores " de réussite), et il faut également être vigilent sur la portée des résultats, comme le montrent Kestemont et al. (2019 : 13) en pointant les limites de certaines classifications (pour eux le deep learning " have so far not led to a major breakthrough in the field "), et en concluant ainsi un état des lieux sur le sujet: " a more promising research direction might be to move away from closed-set classifiers (with a naive reject-option), towards purely open-set classifiers " En nous situant dans la perspective de travaux d'analyse du discours outillée, qui font cohabiter IA et textométrie, nous souhaitons présenter un algorithme de machine learning capable de détecter automatiquement la paternité d'un texte dans un ensemble de corpus de textes/discours d'auteurs connus. La recherche en cours se base sur un corpus hétérogène composé à ce jour de 1000 documents (emails, livres du domaine public, de discours et articles de blog) annotés en fonction de l'auteur, à la suite de travaux exploratoires sur des tweets politiques en contexte électoral (Lam et al. 2021). Nous analysons en pré-traitement plus d'une vingtaine de paramètres différents (distance de Jaccard, fréquence des différentes parties du discours, caractéristiques stylistiques comme la longueur/complexité des phrases, etc., qui ont notamment été identifiés comme pertinents pour cette tâche à partir d'explorations textométriques) que nous passons ensuite dans un algorithme de machine learning (SVM). Les travaux en cours donnent des résultats encourageants, que nous discuterons notamment au regard des résultats donnés par

*Intervenant
†Auteur correspondant: adingbossou@yahoo.fr
‡Auteur correspondant: jeremy.demange.mail@gmail.com
§Auteur correspondant: m.boumrar@hotmail.fr

DeepText. Dans la perspective applicative, nous montrerons que si l'attribution d'auteur est peu utilisée par les juridictions françaises, elle peut s'intégrer aux procédures d'enquête ou de justice.
Brunet, E. (2003). Peut-on mesurer la distance entre deux textes?. *Corpus*, (2).

Brunet, E., Lebart, L. & Vanni, L. (2021). Littérature et intelligence artificielle. Dans Mayaffre D. & Vanni L. (Éds), *L'intelligence artificielle des textes. Des algorithmes à l'interprétation.* 73-130

Kestemont, M., Stamatatos, E., Manjavacas, E., Daelemans, W., Potthast, M., & Stein, B. (2019). Overview of the Cross-domain Authorship Attribution Task at PAN 2019. In *CLEF (Working notes).*
Lam, T., Demange, J., & Longhi, J. (2021). Attribution d'auteur par utilisation des méthodes d'apprentissage profond. Dans *EGC 2021 Atelier" DL for NLP: Deep Learning pour le traitement automatique des langues"*

Mots-Clés: stylométrie, textométrie, machine learning, attribution d'auteur