# Apprentissage actif profond pour le classement de textes en plusieurs classes

Yves Mercadier*, Jérôme Azé*, Sandra Bringay*,**

*LIRMM UMR 5506, Université de Montpellier, CNRS, France
name@lirmm.fr, http://www.lirmm.fr/
**AMIS, Département MIAp, Université Paul-Valéry Montpellier, France

**Résumé.** Récemment, le classement de documents textuels a beaucoup progressé. Cependant, les modèles utilisés doivent généralement s'entraîner au préalable avec de nombreux échantillons étiquetés. Il est possible de diminuer ce nombre d'échantillons en choisissant mieux les données à annoter via des techniques d'apprentissage actif. Cela peut permettre de diminuer le coût du processus en réduisant l'intervention humaine. Dans cette étude, nous adapterons les techniques récentes d'apprentissage actif profond utilisées pour le classement d'images, au cas de l'analyse de textes. En particulier, nous serons attentifs à l'apport de l'apprentissage actif profond selon l'architecture utilisée (LSTM ou CNN). Nous validerons nos hypothèses sur des jeux de données de la littérature.

## 1 Introduction

En classement de textes, la phase d'étiquetage nécessaire à l'apprentissage du classifieur peut s'avérer longue et fastidieuse. Dans ce contexte, l'apprentissage actif, pendant lequel l'oracle intervient pour choisir les exemples à étiqueter, s'avère prometteur. L'intuition est la suivante : en choisissant les exemples intelligemment et non aléatoirement, les modèles devraient s'améliorer avec moins d'effort et donc à moindre coût (c'est-à-dire avec moins d'exemples annotés). Dans cet article, nous conduisons une étude dans l'intention d'évaluer la qualité des processus d'apprentissage actif pour une tâche spécifique de classement multi-classes de textes.

Dernièrement, les réseaux de neurones se sont avérés très efficaces pour le classement de textes, notamment en utilisant des classifieurs de type LSTM (Long Short-Term Memory) et CNN (Convolutional Neural Network). Or, si beaucoup d'approches d'apprentissage actif profond ont été évaluées pour de le classement d'images, à notre connaissance, il n'existe que peu d'études portant sur le texte et sur ce type de classifieur.

Pour cette raison, nous allons évaluer dans cet article les méthodes de réseaux profonds combinées à des approches par apprentissage actif afin de généraliser à de gros volumes de données les connaissances acquises sur un petit échantillon.