Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

La subjectivité dans le journalisme québécois et belge : transfert de connaissance inter-médias et inter-cultures

Article publié dans *Proceedings of the 17th International Conference on Statistical Analysis of Textual Data*

Louis Escouflaire[1, 2] Antonin Descampe[2] Antoine Venant[3] Cédrick Fairon[1]
(1) CENTAL, UCLouvain, Belgique
(2) ORM-EJL, UCLouvain, Belgique
(3) OLST, Université de Montréal, Canada
[louis.escouflaire|antonin.descampe|cedrick.fairon]@uclouvain.be, antoine.venant@umontreal.ca

RESUME

Cet article s'intéresse à la capacité de transfert des modèles de classification de texte dans le domaine journalistique, en particulier pour distinguer les articles d'opinion des articles d'information. A l'ère du numérique et des réseaux sociaux, les distinctions entre ces genres deviennent de plus en plus floues, augmentant l'importance de cette tâche de classification. Un corpus de 80 000 articles de presse provenant de huit médias, quatre québécois et quatre belges francophones, a été constitué. Pour identifier les thèmes des articles, une clusterisation a été appliquée sur les 10 000 articles issus de chaque média, assurant une distribution équilibrée des thèmes entre les deux genres *opinion* et *information*. Les données ont ensuite été utilisées pour entraîner (ou peaufiner) et évaluer deux types de modèles : CamemBERT (Martin et al., 2019), un modèle neuronal pré-entraîné, et un modèle de régression logistique basé sur des traits textuels. Dix versions différentes de chaque modèle sont entraînées : 8 versions 'mono-médias', chacune peaufinée sur l'ensemble d'entraînement du sous-corpus correspondant à un média, et deux versions 'multi-médias', l'une peaufinée sur 8000 articles québécois, l'autre sur les articles belges.
Les résultats montrent que les modèles CamemBERT surpassent significativement les modèles statistiques en termes de capacité de transfert (voir Figures 1 et 2). Les modèles CamemBERT montrent une plus grande exactitude, notamment sur les ensembles de test du même média que celui utilisé pour l'entraînement. Cependant, les modèles entraînés sur Le Journal de Montréal (JDM) sont particulièrement performants même sur d'autres ensembles de test, suggérant une distinction plus claire entre les genres journalistiques dans ce média. Les modèles CamemBERT multi-médias affichent également de bonnes performances. Le modèle québécois notamment obtient les meilleurs résultats en moyenne, indiquant qu'une diversité de sources améliore la généricité du modèle. Les modèles statistiques (mono- et multi-médias) montrent des performances globalement inférieures, avec des variations significatives selon les médias. Les textes québécois sont plus difficiles à classer pour ces modèles, suggérant des différences culturelles dans les pratiques journalistiques entre le Québec et la Belgique. L'analyse des traits révèle que l'importance de certains éléments textuels, comme les points d'exclamation et les marqueurs de temps relatifs, varient considérablement entre les modèles entraînés sur différents médias. Par exemple, les éditoriaux du JDM utilisent fréquemment des points d'exclamation, reflétant un style plus affirmé et polarisant. En revanche, les articles de La Presse présentent des particularités qui compliquent la généralisation de la tâche. En sommme, cette étude démontre la supériorité des modèles neuronaux comme CamemBERT pour la classification de textes journalistiques, notamment grâce à leur capacité de transfert, bien que les modèles basés sur des traits se distinguent par la transparence de leur 'raisonnement'. Elle met également en lumière des différences significatives entre les cultures journalistiques québécoises et belges.

MOTS-CLES : journalisme, grands modèles de langage, comparaison inter-culturelle, transfert de connaissance

31ème Conférence sur le Traitement Automatique des Langues Naturelles,
volume 2 : traductions d'articles publiés, pages 12–13.

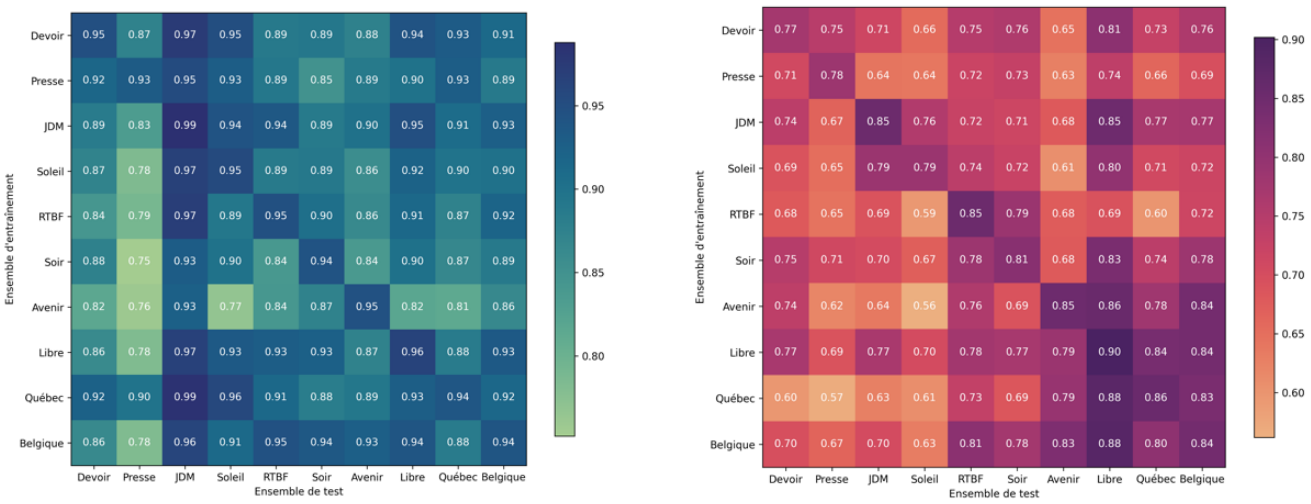

FIGURE 1 & 2 : Exactitudes sur les 10 ensembles de test des 10 modèles CamemBERT (à gauche) et des 10 modèles basés sur des traits (à droite).