LLM TOOL: A HYBRID PIPELINE FOR AUTOMATED LARGE-SCALE TEXT ANNOTATION USING LOCAL LANGUAGE MODELS AND BERT CLASSIFIERS

Auteurs

Lemor, Antoine, Jérémy Gilbert, Shannon Dinan, et Yannick Dufresne

Abstract

The annotation of large-scale text corpora represents a fundamental bottleneck in computational social science research, particularly when dealing with complex multi-label classification tasks in political science. L'annotation de corpus textuels de grande taille constitue un goulot d'étranglement fondamental dans la recherche en sciences sociales computationnelles, en particulier lorsqu'il s'agit de tâches complexes de classification multi-étiquettes en science politique. Nous présentons LLM Tool, une nouvelle chaîne hybride qui combine des modèles de langage locaux de grande taille (LLMs) avec des classifieurs basés sur BERT afin de permettre une annotation entièrement automatisée à grande échelle. Notre approche exploite des LLMs open source de pointe (Gemma3:27B, Llama3.3:42B, Nemotron:42B, DeepSeek-R1:70B, GPT-OSS:120B), exécutés exclusivement sur une infrastructure locale, pour générer des annotations initiales sur des échantillons stratifiés. Ces annotations servent ensuite de données d'entraînement pour des modèles BERT spécialisés capables d'effectuer une inférence efficace à grande échelle. La chaîne met en œuvre une version étendue du schéma de codage du Comparative Agendas Project (CAP), adaptée au discours politique canadien, produisant des annotations structurées selon 21 thèmes de politiques publiques, 9 partis politiques, 2 thèmes spécifiques et 3 dimensions de sentiments. À travers une validation empirique rigoureuse portant sur 1 593 débats parlementaires et articles médiatiques canadiens annotés manuellement, nous montrons que les modèles BERT entraînés à partir d'annotations générées par LLM surpassent largement ceux entraînés sur des annotations humaines, atteignant un score Micro F1 de 0,6673 contre 0,4601 pour les modèles entraînés sur annotations humaines — soit une amélioration de 45 %. Nous validons ce résultat dans trois configurations de jeux de données : Small (1 343 phrases), Large (5 753 phrases) et Extra-Large (12 000 phrases avec isolement complet du test, encore en attente). La configuration Extra-Large, qui garantit une absence totale de contamination entre données d'entraînement et de test, visera prochainement à confirmer la robustesse de notre approche.