Anonymisation réversible des LLM : Une approche basée sur la dispersion des embeddings pour la reconnaissance d'entités nommées

Sébastien Warichet*, Asma Trabelsi* Emmanuel Helbert* Hamed Kallo* Rym Djamouri*

* Alcatel-Lucent Enterprise

1 Introduction

Les grands modèles de langage (LLM) révolutionnent le monde du travail, mais soulèvent des défis majeurs concernant la confidentialité des données sensibles. L'anonymisation réversible offre une solution innovante, permettant de masquer temporairement des informations sensibles tout en préservant la qualité des interactions.

2 Identification et gestion des informations sensibles

L'identification des données sensibles repose sur une reconnaissance d'entités nommées (NER) en trois étapes : (1) les mots monosemiques, simples et non ambigus, sont traités via un dictionnaire, sans calcul inutile ; (2) les mots polysémiques complexes sont isolés et analysés avec un modèle LDA dédié, simplifiant leur classification ; (3) les mots restants sont traités efficacement grâce à la simplification des étapes précédentes.

Pour entraîner nos modèles LDA, nous utilisons une méthode inspirée de la distillation des connaissances : des NER type Flair sont d'abord employés pour labelliser un dataset. Le LDA, léger et rapide à entraîner, garantit un équilibre entre précision et flexibilité. Les entités détectées sont anonymisées via des tags uniques (ex. : *ORG_1*), avec désanonymisation contrôlée par des tables sécurisées.Une description succincte des organisations est ajoutée au contexte afin de ne pas dégrader les performances du LLM

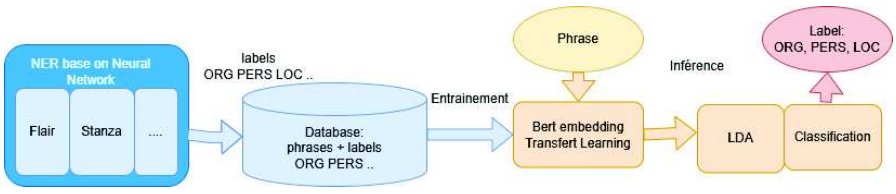

FIG. 1 – *NER description*

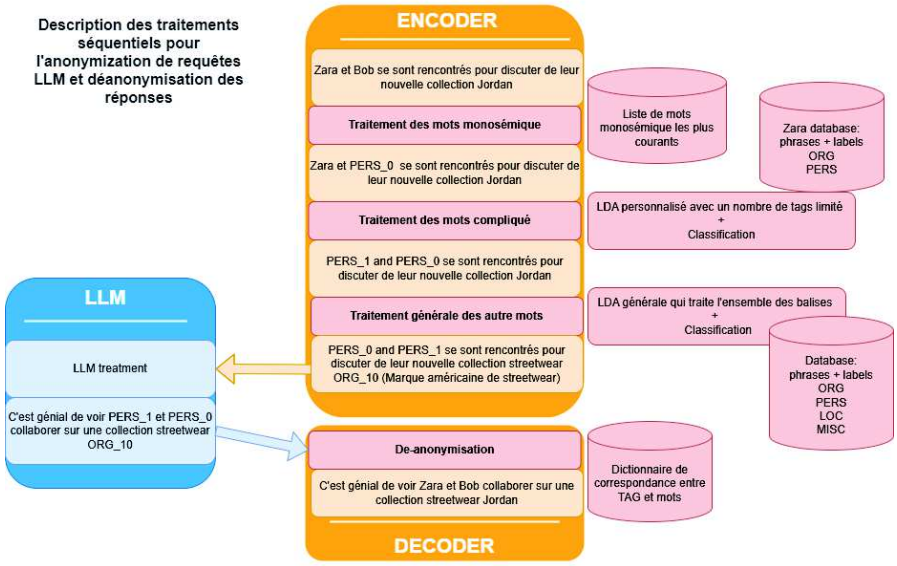

FIG. 2 – *L'anonymisation réversible en 4 étapes*

TAB. 1 – *Comparaison des modèles NER : taille et précision par classe*

Modèle NER	Taille	Précision (ORG)	Précision (MISC)
LDA	7 Ko - 37 Ko	97.93%	**100%**
spaCy	741 Mo	95.56%	82.43%
Stanza	300 Mo	98.80%	84.14%
Flair	1 Go	**98.95%**	99.19%

3 Conclusion

Notre approche d'anonymisation réversible combine la dispersion des embeddings vectoriels et un modèle LDA spécialisé, offrant une solution légère, modulaire et adaptée aux contraintes computationnelles. Contrairement aux réseaux de neurones, qui nécessitent des ressources importantes, notre méthode exploite la simplicité des LDA pour une convergence rapide avec peu de données. En s'appuyant sur les embeddings contextuels générés par BERT, notre solution tire parti du transfert learning pour simplifier l'entraînement et garantir des performances optimales. Idéale pour des applications en temps réel, cette méthodologie allie précision, flexibilité et efficacité.