# GeoNLPlify : Une augmentation spatiale de corpus liés aux crises pour des tâches de classification

Rémy Decoupes[*,***], Mathieu Roche[**,***], Maguelonne Teisseire[*,***]

[*] INRAE, F-34398 Montpellier, France
prenom.nom@inrae.fr,
[**] CIRAD, F-34398 Montpellier, France
prenom.nom@cirad.fr
[***] TETIS, Univ. Montpellier, AgroParisTech, CIRAD,
CNRS, INRAE, Montpellier 34090, France

**Résumé.** L'article *"Deux cygnes retrouvés mort au **Parc de la Tête d'Or** à **Lyon"*** parle-t-il de l'épidémie de grippe aviaire ? Nos travaux proposent d'utiliser l'information spatiale pour générer des données artificielles étiquetées afin d'améliorer les classifications de textes basées sur BERT. Ainsi, après avoir mis en évidence, par des méthodes d'explicabilité, l'importance de l'information spatiale dans les corpus liés à des crises, nous proposons différentes stratégies d'augmentation de données qui tirent profit de ce constat. Notre méthode, GeoNLPlify, est évaluée sur des jeux de données publics (PADI-web et CrisisNLP) et comparée aux augmentations de données classiques.

## 1 Introduction

La dégradation de l'environnement et les effets croissants du changement climatique provoquent une augmentation du nombre de catastrophes et de leurs impacts. Pour permettre une meilleure gestion de ces situations, il devient nécessaire de faire appel à des méthodes d'analyse de données performantes. Le problème auquel nous sommes alors confrontés est le peu de données disponibles. En effet, comme la rareté et la non-similitude des événements sont importants (Buntain et al., 2020), il devient peu pertinent d'appliquer des méthodes d'adaptation de domaine.

En parallèle, le développement des modèles de langue (Large Language Models (LLM)), basés sur les mécanismes d'attention (Vaswani et al., 2017), s'est accru ces dernières années avec des performances exceptionnelles. Même si les LLM sont destinés à être utilisés par transfert d'apprentissage sur des corpus plus petits, ils ont toujours besoin d'un ensemble de données assez important. Différentes techniques d'augmentation de données ont été développées en Traitement Automatique du Langage Naturel (TALN) (Bayer et al., 2022). Leurs objectifs sont d'améliorer les performances d'un modèle de classification de texte en générant artificiellement de nouvelles données étiquetées pour augmenter la taille du corpus d'apprentissage. Cependant, plusieurs approches sont inefficaces quand des LLM sont utilisés car ces derniers sont invariants à certaines transformations (Longpre et al., 2020) telles que le remplacement