Actes de CORIA-TALN-RJCRI-RECITAL 2025
Marseille, France, 30 juin au 4 juillet 2025

Analyse Textuelle et Extraction Géospatiale pour la Surveillance des Crises Alimentaires en Afrique de l'Ouest

Charles Abdoulaye Ngom[1, 2, 3] Maguelonne Teisseire[1, 2] Sarah Valentin[1, 3]
(1) TETIS, Univ. Montpellier, AgroParisTech, CIRAD, CNRS, INRAE, Maison de la Télédetection, 500, rue J.F. Breton, Montpellier, 34090, France.
(2) INRAE, Montpellier, France
(3) CIRAD, Montpellier, France
charles.ngom@inrae.fr, maguelonne.teisseire@inrae.fr, sarah.valentin@cirad.fr

RÉSUMÉ

L'Afrique de l'Ouest fait face à une insécurité alimentaire récurrente, exacerbée par les conflits, le changement climatique et les chocs économiques. La collecte d'informations à une échelle spatio-temporelle appropriée est essentielle au suivi des crises liées à la sécurité alimentaire. Dans ce travail, nous nous intéressons à l'extraction géospatiale à partir de données textuelles, tâche qui s'inscrit dans une approche globale de suivi des crises alimentaires à partir d'articles de presse. Nous évaluons deux modèles d'extraction d'entités spatiales, GLiNER et CamemBERT, en configuration zéro-shot et après ajustement (*fine-tuning*), sur un corpus de 15 000 articles de presse en français couvrant l'actualité du Burkina Faso.

ABSTRACT

Text Analysis and Geospatial Extraction for Food Crisis Monitoring in West Africa

West Africa faces recurrent food insecurity, exacerbated by conflict, climate change and economic shocks. Gathering information at an appropriate spatio-temporal scale is essential for monitoring food security crises. In this work, we focus on geospatial extraction from textual data, a task that is part of a global approach to monitor food crises from news articles. We evaluate two spatial feature extraction models, GLiNER and CamemBERT, in zéro-shot configuration and after adjustment (*fine-tuning*), on a corpus of 15,000 French-language press articles covering events in Burkina Faso.

MOTS-CLÉS : Crise alimentaire, Reconnaissance d'entités spatiales, GliNER, CamemBERT, Afrique de l'Ouest.

KEYWORDS: Food crisis, Spatial Entity Recognition, Spatial Extraction, GliNER, CamemBERT, West Africa.

ARTICLE : **Soumis à** CORIA 2025.

1 Introduction

En 2023, environ 48 millions de personnes en Afrique de l'Ouest et d'Afrique centrale étaient en situation d'insécurité alimentaire aiguë (phases 3 à 5 de IPC), c'est-à-dire incapables de subvenir régulièrement à leurs besoins alimentaires de base (OXFAM, 2022). Ce niveau de faim représente un record en dix ans dans la région. Les projections pour l'année en cours confirment la gravité de la situation. Selon l'analyse du Cadre Harmonisé, 47 millions de personnes au Sahel et en Afrique de l'Ouest nécessiteront une assistance alimentaire et nutritionnelle d'urgence de juin à août

2025 (FCPN, 2025). Les crises alimentaires résultent rarement d'un facteur unique mais plutôt d'une convergence de multiples causes interdépendantes, parmi lesquelles les événements climatiques, les chocs économiques et les conflits figurent parmi les plus importantes (Connolly-Boutin & Smit, 2016; WFP, 2024). Dans ce contexte, la détection précoce et la surveillance des crises alimentaires revêtent d'une grande importance. Plusieurs études ont démontré l'intérêt d'exploiter les données textuelles issues des médias pour prédire ou expliquer ces crises (Balashankar *et al.*, 2023; Ahn *et al.*, 2023). L'extraction d'événements ou de facteurs déclencheurs de crises alimentaires à partir de données textuelles des articles de presse consiste à identifier leur nature (e.g. augmentation des prix, problème de production agricole, etc.) et à les localiser dans le temps et l'espace. Dans nos travaux, nous développons une approche globale de surveillance des crises alimentaires à partir d'articles de presse. Notre approche suit une méthodologie par pipeline : nous extrayons d'abord les entités spatiales et temporelles, puis nous identifions l'événement défini par les relations entre ces entités. Nous ne privilégions pas une approche conjointe (extraction simultanée des entités et de leurs relations), car nous ne disposons pas de données annotées (Zhao *et al.*, 2024).

Dans cet article, nous nous intéressons à la première étape de cette pipeline, à savoir l'extraction d'entités spatiales (noms de pays, régions, provinces, villes ou villages). La localisation joue un rôle central : identifier où une crise alimentaire est susceptible d'être déclenchée est indispensable pour évaluer son impact et organiser une réponse appropriée. L'extraction des entités spatiales permet ainsi de géoréférencer les crises et de faciliter la cartographie des zones affectées (Zhou *et al.*, 2023). Plus précisément, nous analysons deux méthodes d'extraction d'entités spatiales dans le contexte géographique du Burkina Faso : GLiNER (Zaratiana *et al.*, 2024) et CamemBERT (Martin *et al.*, 2020). Il s'agit de comparer ces deux modèles afin d'évaluer lequel est le mieux adapté pour détecter les lieux, soit sans aucun affinement ("zéro-shot"), soit après une spécialisation sur des données spécifiques à un contexte géographique donné (fine-tuning). Cette démarche est d'autant plus pertinente que les noms de lieux africains sont souvent sous-représentés dans les ensembles de données d'entraînement des modèles d'extraction d'entités nommées (Named Entity Recognition, ou NER, en anglais) existants, limitant ainsi leur efficacité dans ces contextes spécifiques (Adelani & Abbott, 2021). Nous appliquons notre approche à un corpus d'articles de presse en français couvrant l'Afrique de l'Ouest, et plus particulièrement le Burkina Faso (Deléglise *et al.*, 2022).

Dans la suite de cet article, nous présentons d'abord une revue des travaux connexes portant sur l'analyse des données textuelles dans le contexte des crises alimentaires, puis nous détaillons les approches existantes en reconnaissance d'entités nommées (NER) spécialisées pour les entités spatiales. Nous exposons ensuite notre méthodologie, en particulier les étapes de prétraitement des données, l'extraction des entités sans entrainement préalable sur les entités spatiales du Burkina (mode zéro-shot), et enfin l'ajustement de ces modèles (fine-tuning) selon la hiérarchie spatiale spécifique du pays. Les résultats obtenus sont ensuite présentés, analysés et discutés, mettant en évidence les avantages et limites des deux modèles de reconnaissance d'entités. Enfin, nous concluons sur les implications pratiques de notre étude pour la surveillance des crises alimentaires et proposons des perspectives pour de futurs travaux.

2 Travaux Connexes

Nous abordons d'abord les travaux existants sur l'analyse textuelle appliquée à la surveillance des crises alimentaires, en soulignant leurs forces et limites mais aussi comment l'identification des entités spatiales est prise en compte. Ensuite, nous examinons les approches de reconnaissance d'entités nommées spécifiquement dédiées aux entités spatiales, en comparant notamment les modèles qui seront évalués dans notre étude.

2.1 Analyse de données textuelles appliquées aux crises alimentaires

Ces dernières années, plusieurs travaux se sont intéressés à l'utilisation de données textuelles issues des médias pour anticiper et comprendre les crises alimentaires, en particulier dans des contextes où les indicateurs traditionnels sont souvent insuffisants ou retardés. Dans l'étude de Balashankar et al. (Balashankar *et al.*, 2023), une méthode supervisée est proposée pour extraire des signaux anticipateurs de crises alimentaires à partir d'un vaste corpus d'articles de presse. Les auteurs utilisent un parser sémantique pour identifier des expressions textuelles (*unigrams, bigrams, trigrams*) associées aux crises par des relations causales. Ahn et al. (Ahn *et al.*, 2023) présentent quant à eux un modèle de deep learning multitâche qui exploite exclusivement les données textuelles pour prédire non seulement l'insécurité alimentaire, mais aussi des indicateurs socio-économiques tels que le prix des denrées et l'instabilité sociale. Grâce à des techniques d'encodage basées sur un modèle pré-entraîné (DistilBERT) et un mécanisme d'attention permettant d'extraire des "gists"(segments de texte particulièrement informatifs), le modèle fournit également une interprétation des signaux de crise. Dans les deux articles, l'extraction d'entités spatiales n'est pas abordée car les deux corpus ont été extraits de bases de données existantes d'archives d'actualités (le projet GDELT et FACTIVA) et sont déjà géolocalisées (à l'échelle régionale ou nationale). Dans (van Wanrooij *et al.*, 2024) est proposée une approche non supervisée, fondée sur le topic modeling via BERTopic, pour extraire les thèmes latents au sein d'articles de presse. En appliquant un processus en plusieurs étapes (encodage des articles, réduction de dimensionnalité, clustering par K-means et extraction de mots-clés via TF-IDF), cette méthode permet de générer des signaux temporels qui, une fois agrégés, contribuent à la prédiction de l'insécurité alimentaire. Les auteurs associent chaque article à une localisation la plus fine possible grâce à une modèle de langue génératif (ChatGPT-3.5-turbo), puis géocodent les entités spatiales extraites grâce à l'API Google Maps. Malgré l'excellente performance de l'approche pour la géolocalisation, les auteurs ont constaté que la géolocalisation d'une article à une échelle régionale (sous-natioanle) n'améliorait pas la performance du modèle de prédiction.

Ces travaux mettent en évidence le potentiel prometteur de l'analyse textuelle pour anticiper les crises alimentaires. Toutefois, les approches actuelles montrent certaines limites importantes : elles reposent principalement sur des méthodes supervisées nécessitant de grandes quantités de données annotées ou sur des techniques non supervisées qui peinent à intégrer une granularité spatiale fine. De plus, ces méthodes exploitent souvent des codages géographiques prédéfinis ou sous-estiment l'importance de la localisation précise, notamment à l'échelle de la ville ou du village, ce qui limite leur efficacité dans notre contexte d'étude. Il est donc essentiel de développer des approches capables de mieux capturer les entités spatiales fines afin d'améliorer la précision des modèles de prédiction de crises alimentaires.

2.2 NER pour les entités spatiales

Pour extraire les entités spatiales des articles de presse, nous comparons deux modèles : GLiNER et CamemBERT. Cette comparaison permet de mettre en évidence le compromis entre flexibilité en mode zéro-shot et performance spécialisée sur un corpus en français.

GliNER est un modèle de reconnaissance d'entités qui se démarque par sa capacité à traiter simultanément des entités de types variés, définis sous forme de libellés textuels. Il s'appuie sur un encodeur bidirectionnel de type BERT, un module de représentation de segments et un module de représentation des entités à extraire (Zaratiana *et al.*, 2024). L'un de ses atouts majeurs réside dans sa faculté à reconnaître de nouveaux types d'entités en mode zéro-shot, sans exiger de données annotées spécifiques, grâce à un entraînement large et multi-type (e.g. Pile-NER). Les résultats expérimentaux montrent que GLiNER dépasse, en configuration zéro-shot, des systèmes plus volumineux (ChatGPT

ou d'autres modèles de langage open-source) pour la détection d'entités, tout en offrant une efficacité supérieure à des modèles génératifs séquentiels. Cependant, l'algorithme "gourmand" de décodage peut rater certaines combinaisons optimales lorsque plusieurs entités se chevauchent, ce qui constitue sa principale limitation.

CamemBERT est un modèle de langue monolingue optimisé pour le français, dérivé de la famille RoBERTa (Antoun *et al.*, 2024). Pré-entraîné sur un large corpus francophone (OSCAR), il tire parti de diverses optimisations comme le masquage dynamique et la suppression de la prédiction de phrase suivante, pour mieux modéliser le contexte des mots. Ce modèle obtient régulièrement des performances de pointe en reconnaissance d'entités, surpassant les modèles multilingues (mBERT) et d'autres approches plus anciennes. Sa force réside dans la qualité des représentations linguistiques, particulièrement adaptée aux spécificités orthographiques et lexicales du français. En revanche, ce modèle ne fonctionne pas en mode zéro-shot pour des types d'entités inédits et nécessite un apprentissage supervisé (fine-tuning) sur des données annotées pour chaque nouvelle tâche de NER.

Dans le contexte de l'extraction d'entités spatiales liées aux crises alimentaires en Afrique de l'Ouest, ces deux approches révèlent un compromis important : d'un côté, GLiNER se distingue par sa flexibilité à intégrer de nouveaux types d'entités sans corpus annoté ; de l'autre, CamemBERT offre une précision élevée, mais exige un entraînement supervisé pour chaque nouvelle catégorie. D'autres approches renforcent également l'extraction d'entités spatiales en contexte spécialisé. TopoBERT (Zhou *et al.*, 2023), en particulier, propose un module de reconnaissance de toponymes basé sur fine-tuning de BERT couplée à un réseau convolutif unidimensionnel (CNN1D). Cette méthode conçue pour la reconnaissance de noms de lieux atteint de bonnes performances (f1-score$\geq$0,865). En outre, la résolution des ambiguïtés dans l'identification des toponymes est traitée par des approches de deep learning également, comme celle proposée par une autre étude, qui utilise des architectures neuronales pour améliorer la correspondance des toponymes avec des entités géographiques candidates (Ardanuy *et al.*, 2020).

3 Méthodologie

Nous présentons tout d'abord les données utilisées pour tester nos approches, puis détaillons l'extraction d'entités spatiales en zéro-shot puis en fine-tuning.

3.1 Données

3.1.1 Corpus

Notre corpus d'étude comprend 15 000 articles de presse en français non annotés. Les articles ont été récupérés par web scraping à partir d'une liste de sources en ligne couvrant l'actualité des pays d'Afrique de l'Ouest (par exemple, LeFaso.net, Burkina24). Pour garantir une extraction précise des entités géospatiales, nous avons mis en œuvre une phase de prétraitement visant à optimiser la qualité et la structure des textes. Le nettoyage des articles consiste à éliminer les éléments susceptibles d'introduire du bruit dans l'analyse. Nous supprimons les caractères non désirés en conservant uniquement la ponctuation courante, éliminons les URLs à l'aide d'expressions régulières, et retirons les espaces superflus en début et fin de texte. Considérant les contraintes techniques des modèles d'extraction (fenêtre contextuelle maximale de 384 tokens pour GLiNER et 512 pour CamemBERT), nous segmentons chaque article en sous-textes de 300 mots. Pour préserver la continuité contextuelle, un chevauchement de 20% est maintenu entre segments consécutifs. Cette approche permet aux modèles de traiter l'intégralité du contenu tout en respectant la limite de leur fenêtre contextuelle. Au terme de cette phase, notre corpus se compose de 3445 segments textuels issus du traitement 1000 articles. Après traitement sur l'ensemble du corpus, nous obtenons au final 51 469 segments.

3.1.2 Données de validation

TABLE 1 – Distribution des entités spatiales dans l'ensemble du corpus

Type d'entité	Occurrences
Département	50 125
Pays	33 702
Village	25 310
Région	11 179
Province	4 110

Afin d'évaluer des modèles d'extraction d'entités spatiales, nous avons annoté le corpus de manière automatique en utilisant une liste d'entités spatiales du Burkina Faso. Cette liste est considérée comme notre gold standard. Nous nous appuyons pour cela sur l'« Annuaire statistique 2022 de l'administration du territoire » (Burkina Ministère de l'administration territoriale, 2022), qui décrit la hiérarchie officielle : 7936 villages, 390 départements (dont 351 valides et plusieurs variations orthographiques), 45 provinces, 13 régions, ainsi que diverses formes du nom du pays (six orthographes recensées). Le but est de réduire les erreurs liées aux variantes de nom et de mieux identifier les localités sous-représentées dans les corpus de presse standard. En termes d'occurrences dans notre corpus, comme l'illustre le Tableau 1, les noms des départements et du pays apparaissent le plus fréquemment (50 125 et 33 702 occurrences respectivement), suivis par ceux des villages (25 310 occurrences), des régions (11 179 occurrences) et des provinces (4 110 occurrences).

3.2 Extraction zéro-shot

Cette première étape évalue la capacité des deux modèles de reconnaissance d'entités nommées à détecter des entités spatiales dans notre corpus sans entraînement préalable spécifique. Les segments prétraités sont soumis à GLiNER et CamemBERT pour l'extraction d'entités de type "Lieu". Pour GLiNER, nous exploitons sa capacité zéro-shot en spécifiant le type d'entité sous forme de prompt textuel. En parallèle, nous utilisons une version de CamemBERT préalablement fine-tunée sur la tâche générale de reconnaissance d'entités nommées, notamment les organisations et les lieux (Polle, 2023). Il faut noter que cette version de CamemBERT n'a pas été spécifiquement entraînée sur le contexte géographique du Burkina Faso. Un aspect important de notre approche est l'intégration d'une phase de post-traitement basée sur le géocodage, visant à garantir que seules les entités spatiales appartenant au Burkina Faso soient conservées. Cette étape s'appuie sur la bibliothèque *Nominatim* via l'API d'OpenStreetMap [1] pour vérifier l'appartenance géographique des lieux extraits. Notre processus de filtrage géographique comprend plusieurs étapes. Nous commençons par normaliser les entités détectées en supprimant les espaces et guillemets superflus et en convertissant les chaînes en minuscules. Pour optimiser les performances et respecter les limites d'API, nous implémentons un système de mise en cache des résultats de géocodage. Seules les entités dont le géocode indique une appartenance au Burkina Faso (code pays **'bf'**) et correspondant à des types géographiques pertinents sont retenues. Enfin, nous agrégeons les occurrences pour obtenir une fréquence d'apparition, facilitant l'évaluation comparative des performances des deux modèles. L'évaluation de cette phase zéro-shot porte donc à la fois sur la capacité d'extraction des entités spatiales et sur l'efficacité du processus de géocodage.

3.3 Fine-tuning

La deuxième étape consiste à spécialiser les deux modèles pour la reconnaissance d'entités spatiales propres au Burkina Faso. Pour le fine-tuning, notre corpus est réparti en trois sous-ensembles : 70% pour l'entraînement, 15% pour la validation et 15% pour le test. Le Tableau 2 détaille la distribution des différentes catégories d'entités dans chaque sous-ensemble. Dans cette phase, nous utilisons GLiNER dans sa version multilingue et CamemBERT dans sa version de base (non préalablement fine-tunée sur la tâche NER). Un aspect essentiel de notre protocole d'évaluation concerne la présence d'entités inédites dans l'ensemble de test. Nous avons délibérément inclus dans cet ensemble des

1. https://nominatim.openstreetmap.org/ui/search.html

TABLE 2 – Distribution des entités spatiales dans les ensembles d'entraînement, de validation et de test

	Entraînement		**Validation**		**Test**	
	Occurrences	Entités distinctes	Occurrences	Entités distinctes	Occurrences	Entités distinctes
Pays	23,432	6	5,620	6	4,650	6
Département	34,611	272	8,752	209	6,762	184
Province	2,863	39	706	32	541	35
Région	7,785	13	1,961	13	1,433	12
Village	17,872	1,044	4,201	462	3,237	437

entités spatiales qui n'apparaissent ni dans l'ensemble d'entraînement ni dans celui de validation. Cette configuration permet d'évaluer la capacité des modèles à généraliser et à identifier des entités spatiales nouvelles tout en les classant correctement selon leur niveau hiérarchique. Ainsi, 19,9% des villages, 1,1% des départements et 2,9% des provinces présents dans l'ensemble de test sont totalement nouveaux pour les modèles.

Afin d'affiner efficacement ses capacités de reconnaissance des entités spatiales tout en conservant les connaissances multilingues générales apprises précédemment, le modèle GliNER a été ajusté en réentraînant uniquement les couches responsables des représentations des spans (span_rep_layer) et des couches conçues pour gérer les représentations des types d'entités(prompt_rep_layer), tandis que les autres paramètres issus du modèle pré-entraîné (urchade/gliner_multi-v2.1) ont été gelés. Quant au modèle CamemBERT, les couches d'embeddings ainsi que les premières couches de l'encodeur ont été figées pour conserver les représentations linguistiques initiales du modèle. Seules les couches supérieures de l'encodeur et la couche finale de classification ont été mises à jour, permettant au modèle de focaliser l'apprentissage sur des représentations contextuelles spécifiques aux entités spatiales du Burkina.

4 Résultats

Notre évaluation comparative des modèles GLiNER et CamemBERT pour l'extraction d'entités spatiales du Burkina Faso s'est déroulée selon deux approches : une configuration zéro-shot et une configuration après fine-tuning. Cette section présente les résultats obtenus selon ces deux modalités.

L'évaluation en zéro-shot visait à déterminer la capacité des modèles à identifier des entités spatiales du Burkina Faso sans entraînement spécifique. Cette étape évalue en même temps l'extraction des noms de lieux mais aussi le géocodage qui nous a servi de filtrage pour ne retenir que les entités appartenant au Burkina. Les performances globales des deux modèles sur notre corpus de test sont présentées dans le Tableau 3.

TABLE 3 – Évaluation de l'extraction des entités spatiales des modèles CamemBERT et GliNER, en zéro-shot

CamemBERT			**GliNER**		
Précision	Rappel	F1-score	Précision	Rappel	F1-score
0.82	**0.77**	**0.79**	**0.87**	0.58	0.70

Ces résultats indiquent que GLiNER, malgré une précision inférieure, parvient à identifier une proportion plus importante d'entités spatiales pertinentes sans avoir bénéficié d'un entraînement spécifique au contexte géographique du pays. Bien que la précision des modèles soit satisfaisante, les rappels restent limités (0,58 pour GLiNER et 0,77 pour CamemBERT), ce qui traduit une difficulté des modèles à généraliser en l'absence d'un entraînement spécifique. Cette performance en terme de rappel modérée peut s'expliquer par le fait que les modèles pré-entraînés sur des corpus majoritairement anglophones ou généraux sont moins efficaces lorsqu'ils traitent des toponymes spécifiques à l'Afrique de l'Ouest, souvent sous-représentés dans ces ensembles d'entraînement initiaux.

TABLE 4 – Résultats du fine-tuning de CamemBERT

Type d'entité	**Précision**	**Rappel**	**F1-score**
Pays	0.99	1.00	0.99
Département	0.99	0.99	0.99
Province	0.99	0.98	0.99
Région	0.99	0.99	0.99
Village	0.94	0.98	0.96

Contrairement à la phase précédente, nous n'avons pu réaliser le fine-tuning que sur CamemBERT, le modèle GLiNER n'ayant pas convergé avec la même quantité de données annotées. Les résultats obtenus avec CamemBERT après fine-tuning (voir Tableau 4) montrent une très nette amélioration des performances comparativement à sa configuration zéro-shot. Nous privilégions l'utilisation de la métrique micro-avg en raison du déséquilibre marqué dans nos données entre les différentes classes d'entités spatiales. On note une très haute précision et un rappel presque parfait pour les entités spatiales les plus courantes telles que les pays, régions et départements, avec des scores proches ou égaux à 0.98. Même pour les entités moins fréquentes et souvent plus ambiguës telles que les villages, les résultats restent solides avec une précision et un rappel autour de 0.94 et 0.98 respectivement. La catégorie "village" obtient des résultats légèrement moins élevés, avec une précision de 0,94, un rappel de 0,98 et un score F1 de 0,96. Cette différence peut s'expliquer : le nombre plus important d'entités distinctes dans cette catégorie (plus de 1000 villages dans l'ensemble d'entraînement), la présence significative d'entités inédites dans l'ensemble de test (19,9% des villages).

5 Discussion

La présente étude utilise OpenStreetMap pour le géocodage des toponymes extraits. Ce choix, bien adapté à des prototypes de portée restreinte, présente néanmoins deux limites majeures : une cadence maximale d'environ 1 requête s^{-1} et un seuil quotidien de ≈ 2500 requêtes $jour^{-1}$, avec un risque potentiel de bannissement en cas de sollicitation excessive. Afin de disposer d'une hiérarchie administrative cohérente et conforme à l'organisation territoriale du Burkina Faso, nous avons privilégié l'utilisation d'un document officiel à jour. Cette approche permet de mieux structurer les entités spatiales extraites, notamment à une granularité fine, et d'élargir la couverture des localités non référencées dans l'API.

Notre analyse comparative de GLiNER et CamemBERT, pour l'extraction d'entités géospatiales dans le contexte spécifique du Burkina Faso, permet d'identifier des facteurs déterminants d'efficacité pour les systèmes de surveillance basés sur l'analyse textuelle, notamment la granularité spatiale. Notre

première évaluation, menée sur 1000 articles, visait à comparer les capacités intrinsèques des deux modèles pour l'extraction de noms de lieu du Burkina. La précision élevée de GLiNER témoigne de sa robustesse face à des toponymes peu représentés dans les corpus standards, grâce à son architecture qui combine un encodeur bidirectionnel avec des modules spécialisés de représentation de segments et d'entités. Son entraînement sur des corpus multi-types comme Pile-NER lui confère également une solide capacité à distinguer les entités spatiales d'autres types d'entités. Cependant, GLiNER présente un rappel nettement inférieur (0,58 contre 0,77 pour CamemBERT), indiquant par le fait que GLiNER n'est pas spécifiquement optimisé pour le français, contrairement à CamemBERT, conçu et pré-entraîné exclusivement sur des corpus francophones.

Les résultats de la seconde évaluation de l'étape du fine-tuning révèlent une nette supériorité de CamemBERT, qui parvient à atteindre des performances quasi parfaites pour plusieurs catégories d'entités spatiales. Pour les pays, régions, provinces et départements, le fait que les noms varient peu grammaticalement favorise l'apprentissage et facilite la reconnaissance. Pour les villages, dont la variabilité orthographique et la multiplicité sont plus élevées, la performance diminue légèrement (F1$\approx$0,96). Ici, la difficulté réside notamment dans la présence de nombreuses entités inédites en phase de test (près de 20%), ainsi que dans les variantes d'écriture. Nos tentatives pour spécialiser GLiNER dans les mêmes conditions n'ont pas abouti à une convergence satisfaisante. Plusieurs facteurs peuvent expliquer ce résultat. L'algorithme employé pour le décodage peut rencontrer des difficultés à gérer les chevauchements ou variations d'orthographe courantes pour les noms de lieux, en particulier lorsque les variantes sont nombreuses (cas des villages). De plus, la non-spécialisation en français (GLiNER ayant été principalement entraîné sur des corpus multilingues) affecte la qualité des représentations et rend plus délicate l'adaptation au contexte géographique et linguistique spécifique du Burkina Faso. L'intégration de la hiérarchie administrative du Burkina Faso dans notre modèle permet d'associer automatiquement chaque entité à son niveau administratif (village, département, province, région), créant un contexte géographique structuré pour l'interprétation des alertes. Cette approche facilite également l'analyse à différentes échelles spatiales, essentielle pour comprendre la distribution des facteurs de crise. Pour détecter rapidement les crises alimentaires, qui exigent de repérer précisément les lieux à une échelle spatiale la plus fine possibile, un modèle comme CamemBERT, adapté à la hiérarchie spatiale du pays, s'avère indispensable. Malgré les résultats prometteurs obtenus, notre approche présente plusieurs limites qui ouvrent des perspectives d'amélioration. D'une part, la focalisation sur un contexte géographique restreint au Burkina Faso limite la généralisation des modèles à d'autres régions où la variabilité des toponymes et des pratiques orthographiques pourrait être encore plus marquée. De plus, le recours à une phase de post-traitement basée sur le géocodage, bien qu'efficace pour filtrer les entités non pertinentes, dépend de la qualité et de la couverture des données issues d'OpenStreetMap, ce qui peut induire des erreurs de filtrage.

6 Conclusion

Nous proposons un pipeline d'extraction spatiale qui repère les crises alimentaires dans la presse jusqu'au niveau village, malgré la rareté des toponymes africains dans les corpus NER. CamemBERT, adapté au français et enrichi de la hiérarchie administrative burkinabè, surclasse GLiNER et comble l'échelle trop grossière des outils actuels. Sa performance reste tributaire du géocodage et de la variabilité orthographique. Nous visons à étendre la méthode à d'autres régions, mieux gérer ces variantes et extraire simultanément lieux et dates ; couplée à des LLM, cette approche nourrira des graphes spatio-temporels pour anticiper les crises.

Références

ADELANI D. I. & ABBOTT E. A. (2021). MasakhaNER : Named entity recognition for african languages. *Transactions of the Association for Computational Linguistics*, **9**, 1116–1131. DOI : 10.1162/tacl_a_00416.

AHN Y., YAN M., LIN Y.-R. & WANG Z. (2023). HungerGist : An interpretable predictive model for food insecurity. *arXiv*, (arXiv :2311.10953). DOI : 10.48550/arXiv.2311.10953.

ANTOUN W., KULUMBA F., TOUCHENT R., CLERGERIE D. L., SAGOT B. & SEDDAH D. (2024). CamemBERT 2.0 : A smarter french language model aged to perfection. DOI : 10.48550/arXiv.2411.08868.

ARDANUY M. C., HOSSEINI K., MCDONOUGH K., KRAUSE A., STRIEN D. V. & NANNI F. (2020). A deep learning approach to geographical candidate selection through toponym matching. DOI : 10.48550/arXiv.2009.08114.

BALASHANKAR A., SUBRAMANIAN L. & FRAIBERGER S. P. (2023). Predicting food crises using news streams. *Science Advances*, **9**(9). DOI : 10.1126/sciadv.abm3449.

BURKINA MINISTÈRE DE L'ADMINISTRATION TERRITORIALE D. L. D. E. D. L. S. (2022). Annuaire statistique 2022 de l'administration du territoire. Disponible à : https://www.insd.bf/sites/default/files/2023-12/Annuaire_statistique_national_2022.pdf.

CONNOLLY-BOUTIN L. & SMIT B. (2016). Climate change, food security, and livelihoods in sub-saharan africa. *Regional Environmental Change*, **16**(2), 385–399. DOI : 10.1007/s10113-015-0761-x.

DELÉGLISE H., BÉGUÉ A., INTERDONATO R., D'HÔTEL E. M., ROCHE M. & TEISSEIRE M. (2022). Mining news articles dealing with food security. In M. CECI, S. FLESCA, E. MASCIARI, G. MANCO & Z. W. RAŚ, Éds., *Foundations of Intelligent Systems*, volume 13515, p. 63–73. Springer International Publishing. Series Title : Lecture Notes in Computer Science, DOI : 10.1007/978-3-031-16564-1_7.

FCPN (2025). Food Crisis Prevention Network. Disponible à : https://www.food-security.net/post/situation-alimentaire-et-nutritionnelle-2025.

MARTIN L., MULLER B., ORTIZ SUÁREZ P. J., DUPONT Y., ROMARY L., VILLEMONTE DE LA CLERGERIE É., SEDDAH D. & SAGOT B. (2020). CamemBERT : a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 7203–7219, Online. DOI : 10.18653/v1/2020.acl-main.645.

OXFAM (2022). Crise alimentaire en Afrique de l'Ouest : la pire situation en dix ans, avec 27 millions de personnes souffrant de la faim. Disponible à : https://www.oxfam.org/fr/communiques-presse/crise-alimentaire-en-afrique-de-louest.

POLLE J. B. (2023). camemBERT-ner : model fine-tuned from camemBERT for NER task. Disponible à : https://huggingface.co/Jean-Baptiste/camembert-ner.

VAN WANROOIJ C., CRUIJSSEN F. & OLIER J. S. (2024). Unsupervised news analysis for enhanced high-frequency food insecurity assessment. *Decision Sciences*, **55**(6), 605–619. DOI : https ://doi.org/10.1111/deci.12653.

WFP (2024). Global report on food crises (GRFC) 2024. https://www.fsinplatform.org/report/global-report-food-crises-2024.

ZARATIANA U., TOMEH N., HOLAT P. & CHARNOIS T. (2024). GLiNER : Generalist model for named entity recognition using bidirectional transformer. In K. DUH, H. GOMEZ & S. BETHARD, Éds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for*

Computational Linguistics : Human Language Technologies (Volume 1 : Long Papers), p. 5364–5376 : Association for Computational Linguistics. DOI : 10.18653/v1/2024.naacl-long.300.

ZHAO X., DENG Y., YANG M., WANG L., ZHANG R., CHENG H., LAM W., SHEN Y. & XU R. (2024). A comprehensive survey on relation extraction : Recent advances and new frontiers. *ACM Comput. Surv.*, **56**(11), 293 :1–293 :39. DOI : 10.1145/3674501.

ZHOU B., ZOU L., HU Y., QIANG Y. & GOLDBERG D. (2023). TopoBERT : Plug and play toponym recognition module harnessing fine-tuned BERT. DOI : 10.48550/arXiv.2301.13631.