Revue de la litérature

CAROT Axel, ARISOY Ivan Can,
DARDE Guilhem, NDJINGA NDJINGA Anta Claude

6 décembre 2023

1 Contexte global du projet

L'Afrique de l'Ouest est une région dynamique par sa diversité culturelle et ses richesses naturelles mais elle fait face à des défis importants en matière de sécurité alimentaire. La région, confrontée à une forte croissance démographique et à des conditions climatiques variables, est particulièrement vulnérable aux fluctuations de la production agricole et aux crises alimentaires.

La gestion des risques liés à la sécurité alimentaire émerge comme un enjeu majeur pour le développement durable de l'Afrique de l'Ouest. Les crises alimentaires, notamment les grandes sécheresses des années 70, ont montré la nécessité de mettre en place des systèmes d'alerte précoce efficaces pour anticiper et gérer ces risques. C'est dans ce contexte que s'inscrit notre projet.

Pour aborder ces problèmes, le CIRAD utilise une approche pluridisciplinaire, en intégrant des technologies de pointe et des méthodes d'analyse de données. Ils utilisent, par exemple, des données satellitaires pour suivre les anomalies d'indices de végétation, un indicateur clé de la santé des cultures et de la sécurité alimentaire.

Cependant, une partie de l'information reste inexploitée dans les données textuelles. Notre objectif est donc d'utiliser des techniques de fouille de texte et de traitement automatique du langage naturel (TALN) pour exploiter un corpus de données textuelles sur la sécurité alimentaire en Afrique de l'Ouest. Ce corpus inclut des articles de journaux locaux et des transcriptions de vidéos, principalement en français.

2 Documentation fournie par les commanditaires

Les commanditaires nous ont fourni une documentation composée en partie de 3 articles. (disponibles sur repository github)

2.1 " Explaining food security warning signals with YouTube transcriptions and local news articles "

Ce projet se concentre sur une analyse spatio-temporelle de données textuelles (transcriptions YouTube et articles de journaux locaux) pour comprendre les situations d'insécurité alimentaire. En utilisant des techniques avancées de

fouille de texte sur un corpus de documents en français, l'étude propose un indicateur de sécurité alimentaire basé sur les données textuelles, nommé TXT-FS.

Le document détaille la méthodologie utilisée comprenant l'acquisition de données, le traitement de texte, la sélection d'articles pertinents via l'embedding de mots, et la modélisation de sujets. L'analyse se concentre sur les zones de l'Afrique de l'Ouest, en particulier le Burkina Faso. Le projet met en œuvre des techniques avancées de fouille de texte, comme le Word2Vec et le modèle LDA (Latent Dirichlet Allocation), pour analyser les situations de sécurité alimentaire et développer l'indicateur TXT-FS. Les résultats montrent que cette approche fournit des informations qualitatives significatives et complémentaires sur la sécurité alimentaire et ses caractéristiques spatio-temporelles.

2.2 " How can text mining improve the explainability of food security situations ? "

Basé sur un corpus d'articles de journaux locaux, l'étude propose un pipeline combinant différentes approches d'analyse textuelle pour créer un modèle explicatif évalué sur des données à grande échelle. Ils utilisent des techniques de text mining pour extraire et analyser des informations qualitatives utilisées comme référence de la situation alimentaire nationale et régionale. Le document présente une analyse spatio-temporelle basée sur des textes en français et l'extension de mesures de discrimination pour traiter des données spatio-temporelles.

2.3 " Mining News Articles Dealing with Food Security "

Le document se concentre sur l'utilisation de méthodes de text mining pour analyser la sécurité alimentaire en Afrique de l'Ouest en combinant différentes techniques d'analyse textuelle. L'objectif est d'obtenir un modèle explicatif évalué sur des données à grande échelle.

Leur méthodologie inclut l'utilisation de Word2vec pour le word embedding, VADER pour l'analyse de sentiment, et tf-idf pour l'évaluation de l'importance des termes. Ils ont créé un pipeline pour effectuer une analyse spatio-temporelle de la sécurité alimentaire basée sur les données textuels. Leur étude a permis d'extraire et d'analyser des informations qualitatives utilisées comme référence pour la situation alimentaire nationale et régionale.

3 Approches utilisées et améliorations

3.1 Approches utilisées par les commanditaires

Les documents vus précedemment se concentrent sur l'analyse spatiale et temporelle de données textuelles comme les transcriptions YouTube et les articles de presse locaux. Cette approche permet de capturer des informations

qualitatives sur la sécurité alimentaire dans différentes régions.

Nous avons pu comprendre le développement d'un indicateur innovant basé sur l'analyse textuelle pour évaluer la sécurité alimentaire. Cette méthode offre une perspective complémentaire aux indicateurs traditionnels qui se fondent sur les enquêtes et l'imagerie satellite.

3.2 Propositions d'amélioration

Les documents dont nous disponsons, bien qu'ils intègrent des analyses avancées de données textuelles pour comprendre la sécurité alimentaire, n'exploitent pas les capacités des modèles linguistiques modernes comme BERT et BERTtopic.

De plus, notre corpus de données, principalement constitué de transcriptions YouTube et d'articles de presse, pourrait bénéficier de l'inclusion de données supplémentaires, telles que les transcriptions des émissions de radio, pour une compréhension plus complète des problématiques de sécurité alimentaire.

4 Litérature complémentaire

Dans le cadre de notre recherche sur l'utilisation des modèles BERT, nous avons exploré diverses études et publications pour enrichir notre compréhension et approfondir notre méthodologie.

4.1 " Food Safety Awareness and Opinions in China : A Social Network Analysis Approach " and " Food Quality and Preference "

Nous avons découvert dans les références de cette cette étude, une publication intéressante : "Food Quality and Preference", qui se concentre sur l'utilisation de modèles BERT dans un contexte différent mais applicable.

"Food Quality and Preference" offre un aperçu de l'utilisation efficace des modèles BERT dans l'analyse des données de médias sociaux, en se concentrant sur les perceptions publiques de la viande alternative en Chine. Cette approche est particulièrement pertinente pour notre projet car elle démontre comment les modèles BERT peuvent être appliqués pour analyser des données textuelles complexes et extraire des informations significatives.

Méthodologies utilisées :
Modèles BERT : Cette approche est bénéfique pour notre projet car elle permet une compréhension profonde des données textuelles, ce qui est essentiel pour analyser les discussions et les rapports sur la sécurité alimentaire.

Analyse des Médias Sociaux : L'utilisation de données issues des médias sociaux offre une perspective directe et actuelle sur les perceptions et attitudes du public. Cela peut être particulièrement utile pour comprendre les tendances et les préoccupations liées à la sécurité alimentaire dans différentes régions.

4.2 Projet utilisant BERT

Nous avons cherché à comprendre comment d'autres personnes ont utilisé des modèles BERT dans des contextes similaires.

Utilisation de BERTopic pour la Classification des Contextes : Un exemple d'utilisation de BERTopic pour le modelage de sujets, comme démontré dans ce projet Kaggle. Cette approche permet de classifier de manière efficace les contextes dans de grands ensembles de données textuelles. L'application de BERTopic offre une compréhension détaillée des différents thèmes et sujets présents dans un corpus de texte, ce qui peut être extrêmement utile pour identifier et classer les discussions relatives à la sécurité alimentaire.

4.3 " Coping with low data availability for social media crisis message categorisation "

Cette thèse de Congcong Wang de l'université de Cornell aillant été publié en mai 2023, elle n'a pas encore été cité par ses pairs. Elle est axée sur la catégorisation des messages de crise dans les médias sociaux en utilisant des techniques de traitement automatique du langage. Elle présente plusieurs aspects et méthodes pertinentes pour notre projet sur l'analyse de données textuelles et la sécurité alimentaire en Afrique de l'Ouest.

Les liens avec notre projet :

Adaptation de domaine en cas de crise : La thèse aborde l'adaptation de domaine, où un modèle de catégorisation est formé sur des données annotées d'événements de crises passées et adapté pour catégoriser des messages d'un événement de crise en cours. Cette approche pourrait être transposée à notre projet en adaptant des modèles pré-entraînés pour analyser des données textuelles liées à la sécurité alimentaire.

Apprentissage avec peu de données : La thèse présente des techniques pour développer un modèle avec une petite quantité de données étiquetées. Ces techniques seraient utiles puisque nous disposons de peu de données annotées sur la sécurité alimentaire.

Classification non supervisée : Un modèle est créé en utilisant uniquement des données non étiquetées, ce qui est essentiel lorsque l'on dispose de peu ou

pas de données annotées. Cette méthode pourrait être appliquée pour classer des données textuelles relatives à la sécurité alimentaire lorsqu'il n'y a pas de données annotées disponibles.

Utilisation de modèles de langage pré-entraînés : Tout comme nous la thèse se base sur des modèles de langage pré-entraînés, tels que BERT pour l'adaptation de domaine et l'amélioration de la performance de catégorisation, leur pipeline pourrait nous inspirer.

Techniques d'augmentation pour l'apprentissage avec peu de données : Les techniques d'augmentation auto-contrôlée et itérative auto-contrôlée sont proposées pour générer des messages de crise de formation de haute qualité avec quelques messages étiqueté. Ces méthodes pourraient être adaptées pour enrichir notre ensemble de données sur la sécurité alimentaire.

Utilisation de données pseudo-étiquetées : La méthode P-ZSC, qui pseudo-étiquette un corpus non étiqueté pour la formation du modèle, démontre une performance supérieure comparée à d'autres approches dans la catégorisation des messages de crise. Cette technique pourrait être utile pour étiqueter de manière semi-automatique les données textuelles de notre corpus.

Les méthodes et approches utilisés dans cette thèse, comme l'adaptation de domaine, l'apprentissage avec peu de données, l'utilisation de modèles Bert et l'augmentation des données, pourraient être extrêmement utiles pour aborder les défis de traitement et d'analyse des données textuelles dans le contexte de notre sujet. Ces méthodes nous permettraient de travailler efficacement avec des volumes variables de données annotées/non annotées/pseudo annoté.

5 Liste de mot clés

Pour nous aider dans nos recherches, nous avons constitué **une liste de mot clés :**

Sécurité alimentaire ; Analyse spatio-temporelle ; Fouille de texte ; Transcriptions YouTube ; Afrique de l'Ouest ; Indicateur TXT-FS ; Word2Vec ; Modèle LDA ; Acquisition de données ; TALN ; Embedding ; Analyse de sentiment ; VADER ; tf-idf ; Text mining ;

Nous avons enrichie cette liste au fur et à mesure de nos recherches.