# Résumé Good it (fr)

CAROT Axel

24 septembre 2023

## 1 Introduction

La sécurité alimentaire est une préoccupation mondiale, mais elle connaît des tendances inquiétantes avec une augmentation de l'insécurité alimentaire. Les Nations Unies ont inclus la lutte contre l'insécurité alimentaire parmi leurs objectifs de développement durable. Les systèmes d'alerte précoce actuels se basent sur des données limitées, ce qui offre une image incomplète de la situation.

L'utilisation de données hétérogènes, comme les données spatiales, géographiques et économiques, peut améliorer la prévision de l'insécurité alimentaire. Les données textuelles, souvent sous-exploitées, peuvent être une source d'information précieuse. Nous proposons une méthodologie d'analyse spatiotemporelle de données textuelles pour expliquer l'insécurité alimentaire. Cette méthodologie est appliquée à la région de l'Afrique de l'Ouest en utilisant des transcriptions vidéo et des articles de presse en français. Les résultats montrent que cette approche offre des informations qualitatives complémentaires sur l'insécurité alimentaire et ses caractéristiques spatiales et temporelles.

## 2 Background

Ce texte explique les concepts clés de l'exploration de texte, en mettant l'accent sur les méthodes de vectorisation, la modélisation de sujets et l'évaluation des modèles de sujets.

Les documents textuels étant généralement non structurés et bruyants par nature, ils nécessitent une prétraitement important. Pour extraire des caractéristiques structurées à partir de ces données, on utilise des méthodes de vectorisation, où chaque document est transformé en un vecteur basé sur une procédure ou un schéma de transformation sélectionné. Deux schémas couramment utilisés sont la fréquence des termes (TF) et la fréquence inverse des documents (TF-IDF). La fréquence des termes représente le nombre de fois qu'un mot apparaît dans un document, tandis que TF-IDF prend en compte la fréquence du terme dans l'ensemble des documents et l'inverse de cette fréquence pour augmenter le poids des mots rares. Les documents vectorisés sont essentiels pour les approches d'exploration de texte telles que les modèles de langage, l'incorporation de mots et la modélisation de sujets.

L'incorporation de mots est une méthode pour améliorer la représentation vectorielle des mots. Les méthodes de pointe, comme Word2Vec, génèrent une représentation vectorielle pour chaque mot de la collection de documents. Cette représentation permet de capturer la similarité sémantique entre les mots, et les documents peuvent être représentés en moyennant les vecteurs de leurs mots, ce qui facilite la mesure de similarité entre les documents.

La modélisation de sujets se réfère à des techniques visant à découvrir des thèmes latents dans une collection de documents, ainsi qu'à estimer la probabilité qu'un document appartienne à un certain thème. Latent Dirichlet Allocation (LDA) est une méthode populaire pour la modélisation de sujets, où chaque document est représenté comme une distribution sur un nombre donné de thèmes, et chaque mot est associé à des probabilités d'appartenir à chaque thème. Les mots représentatifs d'un thème peuvent être identifiés en fonction de leurs scores de probabilité. Une fois le modèle LDA entraîné, il peut être utilisé pour inférer la distribution des thèmes dans une collection de documents.

Les modèles de sujets peuvent être évalués à l'aide de différentes mesures, notamment des mesures de cohérence qui évaluent la qualité des thèmes en fonction de la similarité sémantique des mots clés les décrivant.

En résumé, le texte explique les étapes essentielles de la prétraitement des documents textuels, de la vectorisation des documents, de l'incorporation de mots, de la modélisation de sujets avec LDA, et de l'évaluation des modèles de sujets.

# 3 METHODOLOGY

Dans cette section, nous présentons la méthodologie proposée pour l'analyse spatio-temporelle des données textuelles qui soutiendra le processus explicatif des situations d'insécurité alimentaire. Le pipeline utilisera les transcriptions YouTube et les articles de journaux locaux pour soutenir les systèmes d'alerte précoce. Un schéma général du pipeline proposé est représenté dans la Figure 1. Nous décrirons les principaux composants et leur interaction dans les sections suivantes.

## 3.1

ectionAcquisition des données

Notre méthodologie nécessite la construction d'un corpus (collection) de documents pertinents pour la sécurité alimentaire. Pour obtenir un tel corpus à partir de sources hétérogènes, nous devons définir une série d'étapes de prétraitement, différentes pour chaque type de média. Pour les articles de presse, la

première étape consiste à acquérir les articles sur le Web. En revanche, pour obtenir des documents à partir de vidéos, nous proposons de nous appuyer sur les transcriptions générées automatiquement par YouTube (pour cette tâche, nous utiliserons une bibliothèque Python exploitant l'API de transcription/sous-titres de YouTube). Les documents ainsi obtenus fournissent une version textuelle de tout discours dans la vidéo.

## 3.2 Traitement du texte

Les principales étapes de prétraitement appliquées aux documents textuels sont résumées dans la Figure 2.

La première étape est la tokenisation, c'est-à-dire l'extraction des mots du texte, une étape clé dans chaque tâche de traitement automatique du langage naturel (NLP). Ensuite, nous nous appuyons sur la lemmatisation, c'est-à-dire le processus qui remplace les mots par leurs lemmes : par exemple, toutes les déclinaisons d'un verbe (par exemple, "est", "était") sont remplacées par la même racine (par exemple, "être"). Ensuite, nous nettoyons davantage les documents en supprimant les mots vides, c'est-à-dire les mots très fréquents qui n'ont pas de signification en eux-mêmes (par exemple, les prépositions). Le dernier prétraitement consiste en la vectorisation, c'est-à-dire la création d'une représentation vectorielle pour chaque document.

Il convient de noter que les transcriptions automatiques de YouTube nécessitent une étape de prétraitement supplémentaire, car elles ne contiennent aucune ponctuation. La ponctuation peut être ajoutée automatiquement pour obtenir une meilleure représentation des documents et améliorer les étapes de traitement du texte. Pour cette tâche, nous nous appuyons sur la bibliothèque Python Punctuator, formée sur Wikipédia française.

## 3.3 Mots clés pertinents pour la sélection d'articles avec l'incorporation de mots

Après avoir prétraité des documents textuels, nous utilisons Word2Vec pour calculer la similarité sémantique entre ces documents et un lexique de sécurité alimentaire. Cette similarité nous permet de filtrer les articles pertinents pour notre analyse. Nous utilisons un seuil de similarité de 0,36, validé avec des experts du domaine.

## 3.4 Modélisation de sujets

Après la première sélection d'articles pertinents, nous utilisons la modélisation de sujets (LDA) pour organiser ces articles en catégories. Nous choisissons

le nombre de catégories (sujets) à l'avance. Ensuite, nous évaluons les modèles de sujets en utilisant des métriques et une vérification manuelle. Les modèles sont inspectés manuellement pour s'assurer qu'ils sont liés à la sécurité alimentaire.

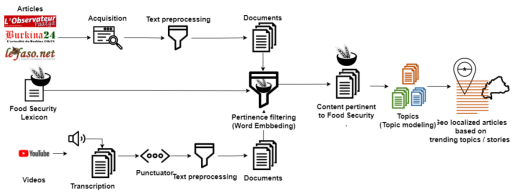



FIGURE 1 – Légende de l'image

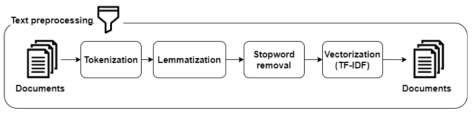



**Figure 2: Text processing**

FIGURE 2 – Légende de l'image

## 3.5 L'indicateur de sécurité alimentaire basé sur le texte (TXT-FS)

Compte tenu de la complexité et de la multifacetté de la sécurité alimentaire, qui est liée à de nombreux facteurs interdépendants, de nombreux indicateurs ont été proposés dans la littérature pour la mesurer (jusqu'à 450 selon [12]). L'utilisation de différents indicateurs en même temps est fondamentale pour évaluer correctement la sécurité alimentaire d'une région donnée, afin de prendre en compte autant d'aspects que possible [3].

Dans ce travail, nous proposons un indicateur géolocalisé de sécurité alimentaire extrait de données textuelles, nommé TXT-FS (indicateur de sécurité alimentaire basé sur le texte), en partant de l'hypothèse qu'il peut servir de proxy efficace pour les indicateurs basés sur des enquêtes. Pour extraire cet indicateur, nous appliquons les étapes suivantes :

Pour chaque document textuel, nous calculons sa similarité sémantique avec un lexique de termes liés aux crises humanitaires et naturelles. Nous utilisons Word2Vec pour calculer la similarité sémantique, tandis que le lexique Crises (CLEX), compilé par des experts du domaine, est disponible en ligne.

Pour chaque document textuel, nous extrayons la mention de la localisation grâce aux techniques de reconnaissance d'entités nommées (NER). Nous utilisons une version du modèle CamemBERT, ajustée pour une tâche NER.

Nous associons chaque document textuel à une province au Burkina Faso en utilisant le géocodeur fourni par la bibliothèque Python GeoPy. Conformément aux pratiques courantes dans la littérature sur la fouille de données d'actualités, seule la première localisation mentionnée dans le texte est utilisée pour extraire la province, en supposant que c'est là que se déroule généralement l'événement discuté dans l'article (c'est-à-dire en cas de mentions multiples de lieux).

Nous agrégeons les informations sur la similarité sémantique au niveau de la province pour obtenir finalement notre indicateur de sécurité alimentaire TXT-FS.

Dans nos expériences, nous calculerons le TXT-FS séparément pour les deux types de documents textuels pris en compte, c'est-à-dire les articles de presse et les transcriptions YouTube. Notez que bien que nous ayons choisi le niveau de la province pour agréger les informations de similarité sémantique en raison de sa pertinence pour les données disponibles et l'échelle nationale de notre analyse, d'autres choix sont possibles, comme des niveaux plus fins (par exemple, les municipalités) ou plus larges (par exemple, les régions).

# 4 RÉSULTATS

Dans cette section, nous montrons l'application du pipeline proposé à deux ensembles de données. Pour les vidéos, un corpus de transcriptions textuelles de vidéos a été extrait des chaînes YouTube de quatre diffuseurs d'actualités (RTB - Radio Télévision du Burkina, Burkina24, ORTB - Office de Radiodiffusion et Télévision du Bénin, TFM - Télé Futurs Medias). Un total de 1109 transcriptions vidéo couvrant la période du 21 janvier 2022 au 12 mars 2022 a été extrait.

Le corpus d'articles de presse est obtenu à partir des versions en ligne de deux journaux locaux du Burkina Faso (Burkina24 et Lefaso.net), pour un total de 22856 articles de presse, relatifs à une période comprise entre 2009 et 2018. Pour l'extraction des articles pertinents, nous nous appuyons sur un lexique de termes liés à la sécurité alimentaire, validé par un panel d'experts du domaine. Nous avons appliqué la modélisation de sujets avec LDA, testant diverses configurations, notamment en variant le nombre de sujets K de 5 à 70, avec une étape de 5, en utilisant deux schémas de vectorisation, TF et TF-IDF. Une inspection qualitative est réalisée pour extraire la configuration la plus efficace.

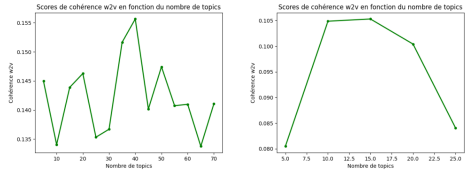
## 4.1 Exploitation de données non conventionnelles

L'analyse des résultats obtenus en utilisant notre pipeline sur des articles de presse montre que le meilleur modèle comporte 40 sujets, ce qui met en évidence des événements liés à la sécurité alimentaire, comme des catastrophes naturelles. La modélisation de sujets permet de détecter des événements non couverts par un lexique préexistant, offrant ainsi une vision plus complète de la sécurité alimentaire.

De même, l'application du pipeline aux vidéos YouTube révèle que le modèle avec 25 sujets est le plus pertinent pour la sécurité alimentaire. Cette approche permet de détecter des événements tels que les impacts du conflit Russie-Ukraine sur la sécurité alimentaire.

## 4.2 Comparaison de TXT-FS avec les indicateurs basés sur des enquêtes

Nous comparons l'indicateur TXT-FS basé sur des données textuelles avec les indicateurs classiques de sécurité alimentaire basés sur des enquêtes. Les résultats montrent que, malgré quelques différences, ces indicateurs présentent des corrélations significatives dans l'évaluation de la sécurité alimentaire des provinces du Burkina Faso. Cela suggère que l'indicateur TXT-FS peut fournir des informations complémentaires aux indicateurs basés sur des enquêtes, couvrant une période plus longue. Cependant, il est important de noter ces différences et de prendre en compte le contexte temporel dans l'analyse.




(a) Coherence of topics from news (b) Coherence of topics from videos

**Figure 3: Coherence evaluation for different number of topics given the best configuration. In (a) coherence values using from 5 to 70 topics on the News articles; In (b) coherence values using from 5 to 25 topics on the video dataset.**

FIGURE 3 – Légende de l'image

**Table 1: Selection of topics related to Food Security, generated from the News articles dataset. In bold, we highlight interesting keywords.**

| Topic ID | Top 10 Keywords | Type |
|---|---|---|
| 14 | **catastrophe** / risque / **urgence** / réduction / national / prévention / action / gestion / **inondation** / **crise** | Event |
| 8 | projet / producteur / **production** / Burkina / produit / **agricole** / local / riz / améliorer / permettre | Agriculture |
| 25 | **campagne** / production / tonne / producteur / **agricole** / ministre / **agriculture** / filière / Burkina / année | Agriculture |
| 28 | **agricole** / **agriculture** / rural / secteur / Burkina / Faso / **développement** / production / ministre / national | Agriculture |
| 5 | pouvoir / santé / **risque** / **aliment** / bon / jour / éviter / devoir / bien / cas | Food Security |
| 35 | **alimentaire** / **sécurité** / nutritionnel / population / situation / **insécurité** / ménage / vulnérable / résilience / Sahel | Food Security |
| 36 | climatique / changement / saison / pluie / **agricole** / **céréale** / pouvoir / vente / Burkina / maïs | Food Security |
| 39 | **crise** / situation / besoin / aide / urgence / **alimentaire** / vivre / vulnérable / million / population | Food Security |
| 26 | pourcent / taux / rapport / **ménage** / résultat / **enquête** / **étude** / niveau / faible / national | Study |
| 29 | enfant / santé / **malnutrition** / région / communautaire / **projet** / **nutrition** / sanitaire / bon / maternel | Study |

FIGURE 4 – Légende de l'image

**Table 2: Selection of topics related to Food Security, generated from the YouTube videos dataset. In bold, we highlight interesting keywords.**

| Topic ID | Top 10 Keywords | Type |
|---|---|---|
| 3 | **ukraine** / africain / aujourd / **russie** / sénégal / **guerre** / pouvoir / **crise** / international / état | Event |
| 11 | aujourd / pouvoir / radio / **céréale** / année / patient / matière / élève / national / pays | Agriculture |
| 12 | **blé** / aujourd / pouvoir / pays / niveau / tonne / état / permettre / politique / beaucoup | Agriculture |

FIGURE 5 – Légende de l'image

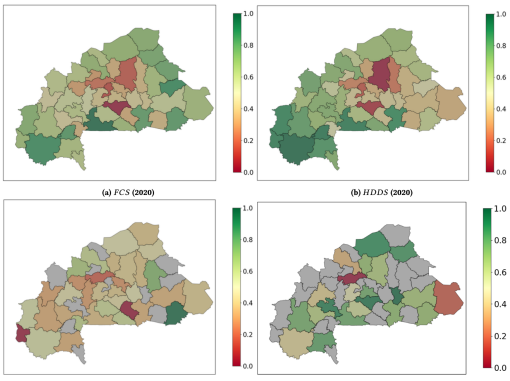
# 5 Conclusion

Dans ce travail, nous avons développé un pipeline d'analyse de données textuelles pour examiner la sécurité alimentaire dans un contexte spatio-temporel. Notre pipeline intègre diverses approches d'analyse textuelle pour créer un modèle explicatif évalué sur des données du monde réel à grande échelle. Nous avons également introduit un nouvel indicateur de sécurité alimentaire basé sur des données textuelles, appelé TXT-FS, en nous concentrant sur la région d'Afrique de l'Ouest, en particulier le Burkina Faso.

Les résultats de notre étude montrent que notre approche offre des informations qualitatives significatives et complémentaires sur la sécurité alimentaire, y compris ses aspects spatiaux et temporels.

Bien que cette recherche représente une avancée importante dans l'utilisation de données textuelles pour l'analyse de la sécurité alimentaire, il reste des possibilités d'amélioration. Par exemple, nous pourrions affiner la façon dont nous associons les lieux à chaque document textuel et utiliser des modèles de langage plus avancés. De plus, nous prévoyons d'incorporer davantage de données temporelles pour mieux comprendre la dynamique de la sécurité alimentaire.

En ce qui concerne les données, nous travaillons à l'élargissement de notre lexique sur les situations de crise et à la collecte de données textuelles provenant de plusieurs pays d'Afrique de l'Ouest. Ces améliorations futures nous aideront à mieux appréhender et à atténuer les problèmes liés à la sécurité alimentaire dans cette région.




**(a)** *FCS* **(2020)** **(b)** *HDDS* **(2020)**

**(c)** *TXT-FS* **on Articles** **(d)** *TXT-FS* **on YouTube transcriptions**

**Figure 4: Food Security levels in the provinces of Burkina Faso measured with different indicators: (a) Food Consumption Score (*FCS*) for the year 2020, (b) Household Dietary Diversity Score (*HDDS*) for the year 2020, (c) *TXT-FS* computed on the news articles corpus, (d) *TXT-FS* computed on the Burkina Faso YouTube transcriptions corpus.**

FIGURE 6 – Légende de l'image