Integrating textual data for enhanced explanation of food crises at subnational scale

Sarah Valentin[1,2,*], Edmond Menya[3,4], Roberto Interdonato[1,2], Mathieu Roche[1,2] and Dickson Owuor[4]

[1]CIRAD, UMR TETIS, F-34398 Montpellier, France.
[2]TETIS, Univ Montpellier, AgroParisTech, CIRAD, CNRS, INRAE, Montpellier, France.
[3]Université de Montpellier, Montpellier, France
[4]Strathmore University, Nairobi, Kenya

Abstract

In an attempt to anticipate Food Security (FS) crises and overcome the limits of existing early warning systems, predictive models can forecast risk indices by combining heterogeneous data. While using different data sources (e.g., satellite imagery, agroclimatic data, food prices) allows to consider various factors that may impact food crises, the explainability of these models remains challenging. In this work, we propose a Food Security indicator solely based on textual data, discerning among different triggers and accounting for possible biases in the spatial coverage of news. We evaluate our approach on a corpus of French-language documents from Burkina Faso and highlight its significance, paving the way for more open and explainable data sources for monitoring food insecurity.

Keywords

Food security, Explainability, Text mining

1. Introduction

Monitoring food security status at an appropriate spatio-temporal level is crucial to detecting early deterioration of food availability and accessibility. Significant progress has been achieved by standardizing methods and indicators to quantify food insecurity [1]. In parallel, the increasing availability of open-access online databases has fostered the development of data-driven approaches that combine heterogeneous data sources [2, 3]. Such predictive approaches utilize secondary data available at specific time intervals and spatial resolutions, often with high precision. Their predictions stem from models based on machine and deep learning, which means they may not be easily interpretable. Other communication channels, such as online articles, provide real-time access to news updates from a specific area. They represent a relevant added value of explainability in predictive models for monitoring crises [4, 5] or to improve predictive models [6, 7].

In this work, we propose an analysis of the explanatory capacity of textual data. We overcome different limits of the first version of a Food Security indicator fully based on textual data, namely *TXT-FS* [4], by improving the data filtering, the type of output and by taking better account of the coverage bias. We evaluate our framework on a corpus of local news from Burkina Faso.

2. Proposed Methodology

Online news are retrieved by scraping web pages from a list of online sources covering the actuality of Western African countries (e.g. LeFaso.net, Burkina24). We applied two filters to ensure the selection

Discovery Science - Late Breaking Contributions 2024

[*]Corresponding author.

sarah.valentin@cirad.fr (S. Valentin); emenya@strathmore.edu (E. Menya); roberto.interdonato@cirad.fr (R. Interdonato); mathieu.roche@cirad.fr (M. Roche); dowuor@strathmore.edu (D. Owuor)
0000-0002-0877-7063 (S. Valentin); 0000-0001-7408-1710 (E. Menya); 0000-0002-0536-6277 (R. Interdonato); 0000-0003-3272-8568 (M. Roche); 0000-0002-0968-5742 (D. Owuor)

of a corpus with a relevant spatial and thematic focus. First, we extracted all spatial entities using a pre-trained transformer-based model dedicated to French textual data, camemBERT [8] and linked each spatial entity into geographic identifiers (e.g., normalized location name and spatial coordinates), relying on the library *GeoPy* and the *GeoNames* geographical database. We selected the online news containing at least one location from Burkina Faso. Second, we adopted a supervised classification approach to filter out the online news not related to food security. We annotated a random sample of the corpus (n=1132) obtaining 61.5% (n=696) of irrelevant news, 18.6% (n= 211) of nearly relevant news (providing general information about food security or information about the consequences of a crisis) and 19.9% (n=225) of relevant news (directly linked with a food crisis). We trained a classifier using a CamemBERT model[1] and used it to predict the relevance category of the remaining online news from the corpus. Eventually, we used the extracted locations to map each online news to one or more provinces of Burkina Faso. This mapping excludes online news in which only large-scale locations are extracted (e.g., regions, country names, etc.).

As food crises are multifactorial, approaches to explicability must make it possible to distinguish between the different drivers (or triggers). To identify them in the text, we adapted an expert-built vocabulary about food security [9] to create a list of food insecurity triggers, classified into 5 categories: extreme weather events, economic shocks and price instability, conflicts, land-related issues and decrease in agricultural production, and a list of crisis-related terms (e.g. famine, hunger). While some keywords carry a negative polarity intrinsically, such as "*drought*", other terms are polarised using quantifier words (e.g., *a drop in agricultural production*). We defined an alert trigger as either (1) the mention of a negative trigger or (2) the mention of a neutral trigger, that is associated with a specific modifier. To detect if a neutral trigger was associated with a relevant modifier, we used the dependency parser dedicated to French data from the library *spaCy* [10].

The food insecurity indicator $TXT - FS_{t,p}$ for a category of triggers t in the province p is:

$$TXT - FS_{t,p} = 1 - \frac{1}{TI_{t,p} + 1} \tag{1}$$

where

$$TI_{t,p} = I_{t,p} + N_{t,p} \tag{2}$$

$N_{t,p}$ represents the number of news articles assigned to the province p in which a crisis trigger was detected. $I_{t,p}$ is 1 if at least one trigger alert was detected in the news assigned to the province p, and 0 otherwise. We added this binary variable to account for the decorrelation between some provinces' food security situation and media coverage, where an approach based solely on the number of articles would risk biasing the results.

The final text-based food insecurity indicator in province p is defined as the weighted average of the food insecurity indicators of each category of triggers:

$$TXT - FS_p = \frac{\sum_{t=1}^{T} TXT - FS_{t,p} \cdot w_t}{\sum_{t=1}^{T} w_t} + \begin{cases} W & \text{if } \forall t, w_t \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

where w_t is the weight associated with the category of trigger t, T is the number of categories of triggers, W is a weight, ranging from 0 to 1, that is added if all weights are non-zero for the province. This allows for the consideration of the synergistic effect of the coexistence of these different factors. In the preliminary analysis, we considered an equal contribution from all categories of trigger weight (w_t = 1). The synergy factor W was set to 0.5.

3. Results & Discussion

From the 15.844 online news published by the selected sources in 2022, 1773 contained at least one location associated with Burkina Faso. Among them, 309 online news were considered relevant (225

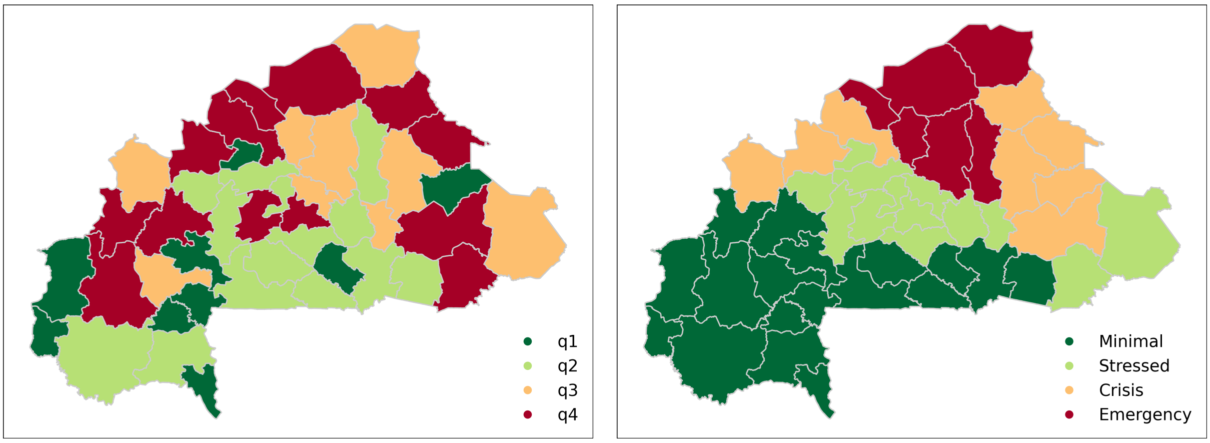
[1] Training parameters: learning_rate=2e-5, num_train_epochs=5, and weight_decay=0.01

were manually labelled to train the classifier, and 84 were automatically labelled). On the training dataset, the classifier obtained an F-score of 0.87 (weighted mean taking into account the class imbalance). The final dataset of online news that were mapped to at least one province consisted of 195 news.

To enable a comparison of $TXT - FS$ with the IPC scores, we split the $TXT - FS$ values into quartiles (Figure 1). The indicator is elevated in several provinces from the Sahelian region, western and eastern provinces and the capital province (Kadiogo). The alert triggers allow us to contextualize the indicator values, especially in the provinces with discrepancies with the IPC value. For instance, the Kompienga province had a high $TXT - FS$ score but was only considered stressed according to the IPC. The corpus contains several references to local situations, in cities such as Pama and Madjoari, particularly affected by the consequences of civil insecurity, as shown in the examples below:

> ”Some populations, particularly in the Burkina Faso towns of **Pama**, Mansila, Kelbo, **Madjoari** and Djibo, are facing dramatic situations.”

This example highlights the importance of considering the different locations mentioned in the articles to enable analysis of local situations. It also points out that a direct comparison between local situation assessment and the IPC, produced on a provincial scale, is not necessarily relevant.



(a) $TXT - FS$, discretized by quantile

(b) IPC (as of 2022-06-01)

Figure 1: Comparison of $TXT - FS$ and IPC values per province. (a) $TXT - FS$ discretized by quantile (2022) (b) IPC (as of 2022-06-01).

In this work, we proposed an indicator that explicitly takes into account the major triggers of food insecurity, enabling the detection of signals even in areas little covered by the press. The resulting maps complement those obtained with household surveys and machine learning-based indicators. This improved indicator contributes to food insecurity early warning and complements existing ones by providing insights into local news.

In the proposed approach, we simplified the temporal aspect by aggregating the data on a yearly scale. However, the driver’s effects on food crises can be long-lasting (e.g. a drought affecting the productive season, conflicts leading to structural production issues, etc.). Assessing the full impact of the triggers may require a better representation of their spatio-temporal footprint, by combining features extracted from the text as well as expert knowledge about the different types of features.

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly for English grammar and spelling check.

References

[1] FEWS NET, What is the IPC?, 2025. URL: https://fews.net/about/integrated-phase-classification[Accessed:Whenever].

[2] E. C. Lentz, H. Michelson, K. Baylis, Y. Zhou, A data-driven approach improves food insecurity crisis prediction, World Development 122 (2019) 399–409. doi:10.1016/j.worlddev.2019.06.008.

[3] P. Foini, M. Tizzoni, G. Martini, D. Paolotti, E. Omodei, On the forecastability of food insecurity, Scientific Reports 13 (2023) 2793. URL: https://www.nature.com/articles/s41598-023-29700-y. doi:10.1038/s41598-023-29700-y, publisher: Nature Publishing Group.

[4] C. T. Ba, C. Choquet, R. Interdonato, M. Roche, Explaining food security warning signals with youtube transcriptions and local news articles, in: GoodIT 2022: ACM International Conference on Information Technology for Social Good, Cyprus, September 7 - 9, 2022, ACM, 2022, pp. 315–322.

[5] H. Deléglise, A. Bégué, R. Interdonato, E. M. d'Hôtel, M. Roche, M. Teisseire, Mining news articles dealing with food security, in: Foundations of Intelligent Systems - 26th International Symposium, ISMIS 2022, Italy, October 3-5, volume 13515 of *Lecture Notes in Computer Science*, 2022, pp. 63–73.

[6] A. Balashankar, L. Subramanian, S. P. Fraiberger, Predicting food crises using news streams, Science Advances 9 (2023) eabm3449. doi:10.1126/sciadv.abm3449.

[7] Y. Ahn, M. Yan, Y.-R. Lin, Z. Wang, HungerGist: An Interpretable Predictive Model for Food Insecurity, 2023. URL: http://arxiv.org/abs/2311.10953.

[8] L. Martin, B. Muller, P. J. Ortiz Suárez, Y. Dupont, L. Romary, É. de la Clergerie, D. Seddah, B. Sagot, CamemBERT: a tasty French language model, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 7203–7219.

[9] H. Deléglise, C. Schaeffer, E. Maître d'Hôtel, A. Bégué, Lexiques en français sur la sécurité alimentaire et les crises (2021). doi:10.18167/DVN1/C5PU01.

[10] ExplosionAI, spaCy · Industrial-strength Natural Language Processing in Python, 2025. URL: https://spacy.io/[Accessed:Whenever].