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

Cheick Tidiane Ba
cheick.ba@unimi.it
Università di Milano, Milan
Milan, Italy

Chloé Choquet
chloe.choquet@cirad.fr
CIRAD, UMR TETIS, Montpellier
Montpellier, France

Roberto Interdonato
roberto.interdonato@cirad.fr
CIRAD, UMR TETIS, Montpellier
Montpellier, France

Mathieu Roche
mathieu.roche@cirad.fr
CIRAD, UMR TETIS, Montpellier
Montpellier, France

ABSTRACT

Food security is a major concern in many countries all over the world. After a relatively long period characterized by a positive trend, the number and severity of food insecurity situations has been growing again in recent years, with alarming projections for the near future. While several Early Warning Systems (EWS) exist to monitor this phenomenon and guide the interventions of governments and ONGs, such systems rely on a narrow set of data types, i.e., mainly satellite imagery and survey data. These data can explain just a limited number of the multiple factors that impact on food security, thus producing an incomplete picture of the real scenario. In this work, we propose a spatio-temporal analysis of unconventional textual data (i.e., YouTube transcriptions and articles from local news papers) to support the explanatory process of food insecurity situations. This data, being completely exogenous to the one used in currently active EWS, can offer a different and complementary perspective on the causes of such crises. We focus on the area of West Africa, which has been at the center of many humanitarian crisis since the beginning of this century. By exploiting state of the art text mining techniques on a corpus of textual documents in French (including video transcriptions extracted from the YouTube channels of four West African news broadcasters and news articles obtained from the online versions of two local newspapers of Burkina Faso) we will analyze food security situations in different regions of the study area in recent years, by also proposing a food security indicator based on textual data, namely *TXT-FS*.

KEYWORDS

food security, text mining, spatiotemporal analysis, topic modeling, social media

ACM Reference Format:
Cheick Tidiane Ba, Chloé Choquet, Roberto Interdonato, and Mathieu Roche. 2022. Explaining food security warning signals with YouTube transcriptions and local news articles. In *Conference on Information Technology for Social Good (GoodIT'22), September 7–9, 2022, Limassol, Cyprus.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3524458.3547240

GoodIT'22, September 7–9, 2022, Limassol, Cyprus
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9284-6/22/09…$15.00
https://doi.org/10.1145/3524458.3547240

1 INTRODUCTION

Food security is known to be a major concern in many countries all over the world. Even if the fight against hunger has never been really close to an end, steady progress was made in the early 2000s, with indicators showing a positive trend in several areas of the world. Nevertheless, recent years have witnessed an inverse trend, with the number and severity of food insecurity situations growing again, with alarming projections for the near future [10]. The importance of this phenomenon is also testified by the fact that the United Nations list the fight against food insecurity as one of the 17 Sustainable Development Goals (SDG 2 - Zero Hunger) to be reached before 2030 in order to "achieve a better and more sustainable future for all"[1]. Multiple and interrelated reasons can be identified for this generalized rise in hunger situations. Just to name a few:

- per capita food availability has been reduced by an increasing number of extreme weather events and by an increasing population growth;
- population displacements due to conflicts often result in a drop in agricultural production and in disorders in the distribution channels;
- structural poverty of populations is aggravated by a difficult global economic context.

Looking at this context, it is easy to see how monitoring food security is a challenging problem that touches several scientific domains, and that needs to be addressed by the use of heterogeneous data from different sources. To monitor, analyze and forecast food insecurity situations at local to sub-national scales, several global food security Early Warning Systems (EWS) have been launched in the past decades, mainly as a reaction to the major droughts happening in West Africa and all the Sahel region during the early 1970s. Such systems, meant to guide the interventions of governments and ONGs, generally rely on a narrow set of data types, since they mainly integrate agroclimatic data from satellite images and indicators extracted from household survey about nutritional, economical and production-related factors. These data can explain

[1] https://www.un.org/sustainabledevelopment/hunger/

just a limited number of the multiple factors that impact on food security, thus producing an incomplete picture of the real scenario: remote sensing data only explains a single face of the phenomenon (i.e., the one regarding agricultural production), and data derived from household surveys is often sparse (due to the fact that their collection is extremely expensive in terms of time and money) and subject to several biases (e.g., related to privacy preservation [22], failure in collecting a complete information [17] and measurement techniques [15]).

On the other hand, it is easy to observe how nowadays great quantities of heterogeneous data are publicly available, that are related at different levels with food security. Some examples may be spatial information (e.g., population density, land use, soil quality), volunteered geographical information (number of hospitals and schools, number and details about violent events) and economic indicators (e.g., price of representing goods). Recent literature has shown how these heterogeneous data can be exploited to predict food security indicators through advanced data science methods [4, 11, 21], e.g., multi-branch neural networks able to integrate data of different types and at different scales, by also taking into account spatial and temporal context. Nevertheless, while the performance of these approaches seem to be promising, they are still far from being optimal, and strongly dependent from the study area taken into account (e.g., availability and quality of the data may not be the same over different country, as well as the correlation with food security indicators). Moreover, like most machine and deep learning approaches, these frameworks act as black boxes, i.e., their goal is to predict food security indicators, but they do not produce any specific information about the causes of food insecurity situations.

Our hypothesis is that textual data can be a valuable source of information in this context, that can be exploited to support the explanation of food insecurity crises in a given area. This data, being completely exogenous to the one used in currently active EWS, can offer a different and complementary perspective on the causes of such crises. Text mining is an extremely active area of data science, that offers nowadays many off-the-shelf solutions to process massive amounts of text, e.g., language models [8], word embedding techniques [23] and many other [1, 9].

At the same time, due to its sparse and noisy nature, such data is rarely exploited along with other heterogeneous data in machine learning frameworks (e.g., like the previously mentioned ones). Anyway, this does not hinder the possibility to exploit it as a complementary source to extract qualitative indicators that are complementary to the quantitative ones.

For this reason, in this work, we propose a pipeline for the spatio-temporal analysis of unconventional textual data to support the explanatory process of food insecurity situations. The proposed pipeline integrates different textual analysis approaches to obtain an explanatory model evaluated on real-world and large-scale data. We focus on the area of West Africa, which has been at the center of many humanitarian crisis since the beginning of this century [10], by studying a corpus of textual documents in French, including textual video transcriptions extracted from the YouTube channels of four news broadcasters (RTB - Radio Télévision du Burkina, Burkina24, ORTB - Office de Radiodiffusion et Télévision du Bénin, TFM - Télé Futurs Medias) and news articles obtained from the online versions of two local newspapers of Burkina Faso (Burkina24 and Lefaso.net). By exploiting state of the art text mining techniques on this corpus we will analyze food security situations in different regions of the country in recent years, by also proposing a food security indicator based on textual data, namely TXT-FS. Note that taking into account a collection in French represents a further challenge, given the reduced number of available resources for this language (e.g., pretrained text mining models). The results of our analyses have proven how our approach provides significant results that offer distinct and complementary qualitative information on the subject of food security and its spatial and temporal characteristics.

2 BACKGROUND

Being generally unstructured and particularly noisy by nature, textual documents require a significant amount of preprocessing. In the following, we will cover the required background on text mining key concepts used in this work.

Vectorization methods. A common way to extract structured features from these data is through vectorization methods, i.e., the creation of a vector for each document, based on a selected transformation procedure or scheme. This is needed because most of the text analysis algorithms are not suited for documents of arbitrary length, and require numerical feature vectors with a fixed size. Two widely used transformation schemes are Term Frequency (TF) and Term Frequency-Inverse Document Frequency (TF-IDF) weighting schemes [18]. Given a corpus of textual documents D, we can define a sorted set W of the terms encountered. Then each document $d_j \in D$, where $j = 1...N$ and $|D| = N$, can be represented by a vector v of size $|W|$: in v, the i-th position of the vector contains the weight of the i-th word in the selected document. In TF, the weight is the term frequency $TF_{i,j}$ which represents the number of times the word w_i occurs in the document d_j. Whereas in TF-IDF, the TF value is multiplied by the Inverse Document Frequency IDF. The document frequency of a word w_i, Df_i is the number of documents w_i appears in, divided by the total number of documents N. Its inverse $1/Df_i$ is used to increase the weights of rare words that are more discriminative in a given set of articles. The use of vectorized documents is crucial for all the main text mining approaches, such as language models, word embedding and topic modeling. Word embedding is a method to improve the vector representation. The state of the art methods, like Word2Vec [23] generate an embedding, i.e., a vector of real values for each of the words in a collection of documents. The vector representation is learnt from text in such a way that words with similar meaning will be represented by similar vectors. Another interesting property of the embeddings generated by these models is that a document can be represented by averaging the vectors of each of its words [16], thus allowing the evaluation of similarity measures between documents. Similarity is usually computed using standard measures for vectors, like cosine similarity [18].

Topic modeling. Topic modeling refers to a family of techniques that aim to discover latent (i.e., not previously defined) topics in a collection of documents, also estimating the probability for a document to belong to a certain topic. The obtained topics allow a better exploration of a document collection, as well as the discovery of keywords that characterize a certain topic. Latent Dirichlet

Allocation (LDA) [1] is the most popular method for topic modeling. LDA is an unsupervised method that is used to extract the latent topics from a document collection, given the number k of topics to extract. The LDA method creates a topic per document model and words per topic model. Given the $n \times d$ document-term matrix, we can represent it as the product of an $n \times k$ matrix U and a $d \times k$ matrix V so that $D = UV^T$. The matrix U is the topic per document model: each document is represented as a distribution on k topics as a k-dimensional vector. The matrix V is the words per topic model: given one of the d words we have a k-dimensional vector of probability scores, one for every topic. The representative words for a topic can be found by sorting the words by their probability score. During the training phase, the LDA model adjusts the probability of words belonging to a topic. This is done by processing each document and randomly assigning each word in the document to one of k topics and adjusting the probability for each word to belong to topic t. Once the model has been trained, it is possible to infer the topics distribution for each document in the collection, as well as for new unseen documents. The produced topic models can be evaluated in many ways. Among the established evaluation metrics, the state of the art offers coherence metrics [25]. Coherence scores provide a quantitative evaluation of the quality of the topics produced, based on the homogeneity (i.e., semantic similarity) of the keywords describing each topic.

3 METHODOLOGY

In this section, we present the proposed pipeline for the spatio-temporal analysis of textual data, that will support the explanatory process of food insecurity situations. The pipeline will leverage YouTube transcriptions and articles from local news papers for the support of Early Warning Systems. A general schema of the proposed pipeline is depicted in Figure 1. We'll describe the main components and how they interact in the following sections.

3.1 Data acquisition

Our methodology requires the construction of a corpus (collection) of documents that is pertinent to food security. To obtain such corpus, from heterogeneous sources, we need to define a series of preprocessing steps, different for each media type. For News articles, the first step is the acquisition of the articles from the Web. Whereas, to obtain documents from Videos, we propose to rely on the transcriptions automatically generated by YouTube (for this task, we will rely on a Python library exploiting the YouTube Transcript/Subtitle API[2]). The obtained documents provide a textual version of all speech in the video.

3.2 Text processing

The main preprocessing steps applied to the textual documents in use are summarized in Fig. 2.

The first step is tokenization, i.e., the act of extracting words from text, a key stage in every NLP (Natural Language Processing) task. Then, we rely on lemmatization [19], i.e. the process that replaces words with lemmas: for example all declination of a verb (e.g., "is", "was") are replaced by the same root (e.g., "be"). Next, we further clean the documents by removing the Stopwords, i.e. very frequent words that do not carry any significance by themselves (e.g., prepositions). The last preprocessing is vectorization, i.e. the creation of a vector representation for each document.

Note that automatic YouTube transcriptions need an extra preprocessing step, since they lack any sort of punctuation. Punctuation can be added automatically to obtain a better representation of the documents and improve the text processing steps. For this task, we rely on the Punctuator Python library, trained on French Wikipedia[3].

3.3 Pertinent keywords for article selection using word embedding

Given a collection of preprocessed textual documents, we need to select a subset of candidate articles related to Food Security (i.e., we need to filter out documents that are off-topic with the goal of our analysis). This is done by means of a semantic similarity computation with respect to a Food Security lexicon, i.e., a collection of terms, provided by domain experts, that are strictly related to food security situations. The lexicon used in our work is freely available online [6]. For semantic similarity, we rely on Word2Vec [23], a well known word embedding model. With the Word2Vec model, we are able to represent every word, both in articles and in food security Lexicon, in the same embedding space. We can treat the lexicon as a document. Hence we can represent both the Lexicon and the other documents as an average of can be represented by an embedding, computed as the average of the words. Therefore, given the embedding of the lexicon d_l and the embedding of the i-th document d_i, we can compute a similarity score $sim(d_l, d_i)$ (in the range $[-1, 1]$) for each textual document. Then a semantic similarity threshold thr_{sim} can be chosen to decide whether a document is pertinent to food security, allowing us to consider only the most pertinent articles from the entire corpus. In this work, we use a threshold of 0.36, that was chosen and validated with the aid of domain experts [5].

3.4 Topic modeling

While in the previous step we performed a first selection of pertinent articles, we want to organize the selected articles in a way that would be easily consulted by a domain expert. With topic modeling, we can find latent topics in the collection of pertinent articles, allowing us to focus on a subset of articles that are more related to food security. Among the topic modeling methods, we selected LDA [1]. With LDA, we need to set a number of topics k in order for the model to be trained to group the articles in the requested number of topics, based on their content. By doing this, we can then perform the selection of the most relevant topics. The obtained topic models are evaluated with a combination of evaluation metrics and with manual verification. To guide the model selection, we compute the well established coherence scores [25]: these scores give a quantitative evaluation of the topics generated, based on the semantic similarity among the keywords describing each topic. Then the topic models are manually inspected to verify the presence of keywords related to Food Security. We opt for manual verification in this case given the fact that we chose relatively low values of k (which makes the manual inspection feasible),

[2] https://github.com/jdepoix/youtube-transcript-api

[3] https://pypi.org/project/punctuator/

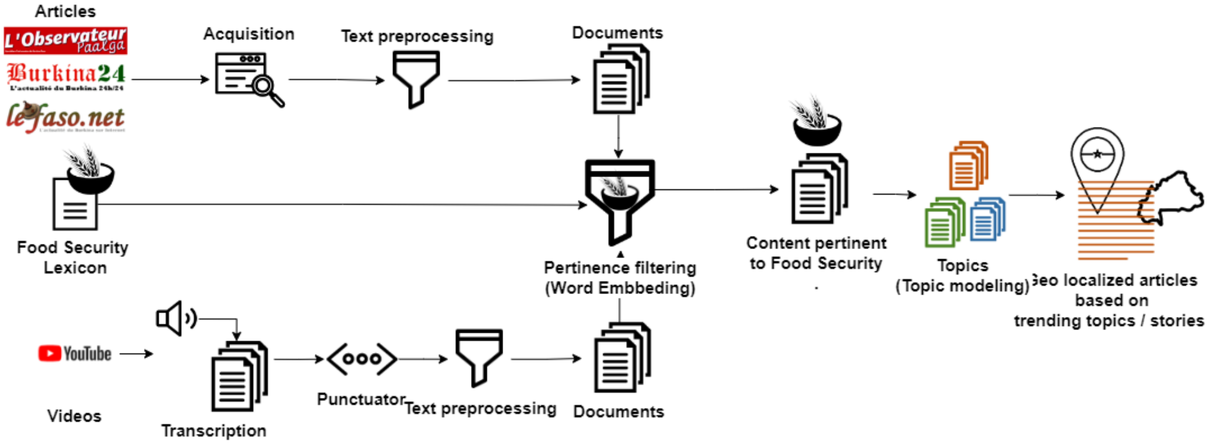


Figure 1: Pipeline: pipeline for the spatio-temporal analysis of textual data to support the explanatory process of food insecurity situations. Videos and news articles are preprocessed to generate a collection of documents. With the support of a lexicon (i.e., a collection of words provided by food security experts) we can filter the most relevant articles for our task. Finally, through topic modeling we can organize documents in topics: the most relevant topics will be used to extract qualititave indicators needed to support the Food Security Early Warning Systems.

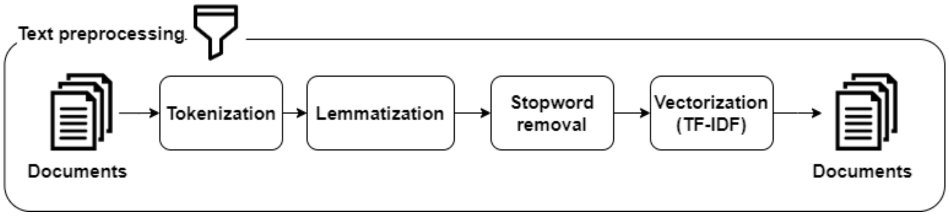


Figure 2: Text processing

and because coherence alone does not give any information about the semantic of the topic (e.g., relevant or not for food security). Anyway, in case of a high number of topics, this step can also be automated through the use of semantic similarity, e.g., by computing the average similarity of the documents belonging to each topic with respect to a domain-specific lexicon, as done in [13].

3.5 The TeXT based Food Security indicator (*TXT-FS*)

Given the fact that food security is an extremely complex and multifaceted concept, related to many interdependent factors, many indicators have been proposed in literature to measure it (up to 450, according to [12]). Using different indicators at the same time is fundamental to properly assess the food security of a given area, in order to take into account as many aspects as possible [3].

In this work, we propose a geolocalized indicator of food security extracted from textual data, namely *TXT-FS* (TeXT based Food Security indicator), with the hypothesis that it can be used as an effective proxy for indicators based on survey data. To extract this indicator, we apply the following steps:

(1) For each textual document, we compute its semantic similarity to a lexicon of terms related to humanitarian and natural crises. Word2vec (w2v) [23] trained on the French Wikipedia is used to compute semantic similarity, while the Crises LEXicon (*CLEX*), compiled by domain experts, is freely available online [6].
(2) For each textual document, we extract the location mentioned in it through Named Entity Recognition (NER) techniques. A version of the CamemBERT [20] model fine tuned for a NER task is used in this work[4].
(3) We associate each textual document to a province in Burkina Faso by using the geocoder provided by the GeoPy Python library[5]. As commonly done in news mining literature, only the first location mentioned in the text is used to extract the province, under the hypothesis that it is the one where the event discussed in the article usually happens (i.e., in the case that multiple locations are mentioned).
(4) We average information about semantic similarity at the province level, to finally obtain our food security indicator *TXT-FS*.

In our experiments, we will compute *TXT-FS* separately for the two types of textual documents taken into account, i.e., news article and YouTube transcriptions. Note that, while we have chosen the province level to aggregate semantic similarity information, for the

[4] https://huggingface.co/Jean-Baptiste/camembert-ner
[5] https://geopy.readthedocs.io/en/stable/

fact that is was the best fit for the available data and the national scale of our analysis, different choices are possible, i.e., more fine grained ones (e.g., municipalities) or coarse ones (e.g., regional).

4 RESULTS

In this section, we show the application of the proposed pipeline to two datasets. For videos, a corpus of textual video transcriptions was extracted from the Youtube channels of four news broadcasters (RTB - Radio Télévision du Burkina, Burkina24, ORTB - Office de Radiodiffusion et Télévision du Bénin, TFM - Télé Futurs Medias). From the four channels, a total of 1109 of video transcriptions was extracted, covering the period from 21 January 2022 to 12 March 2022. The news articles corpus is obtained from the online versions of two local newspapers of Burkina Faso (Burkina24 and Lefaso.net), for a total of 22856 news documents, related to a time period between 2009 and 2018 [7]. For the extraction of pertinent articles, we rely on a lexicon, i.e., a list of terms, related to food security [6]. The terms contained in the lexicon were validated by a panel of domain experts. We applied topic modeling with LDA, testing various configurations: we varied the number of topics K in the range $5 \leq K \leq 70$, with a step of 5, using two vectorization schemes, TF and TF-IDF based. A qualitative inspection is performed for the extraction of the most effective configuration.

4.1 Leveraging unconventional data

We first discuss the results obtained by applying the pipeline on the news articles dataset. The best configuration was obtained using TF based vectorization, and filtering terms with minimum DF of 0.05 and maximum DF of 0.86. Coherence measures shown in figure 3a suggest that the best model is the one with 40 topics and a successive verification confirmed the observation. In Table 1 we present a selection of the most relevant topics for food security.

Topic 14 highlights the importance of the topic modeling phase we propose. In topic 14, the keywords, such as "inondation" (*flood*) "catastrophe"(*catastrophe*) and "urgence" (*urgency*) indicate that there has been a crisis event. To this regard, Burkina Faso has indeed been the target repeated floods in 2009 and 2010 [2, 24]. Note that while the word "inondation" is not included in the Crises LEXicon *CLEX*, so that with topic modeling we are able to detect a potential cause of issues in the region, and at the same time we may select keywords to include in the lexicon for future analyses. Among the other topics related to food security, we can observe that topics 8, 25, 28 feature prominently keywords related to agriculture: in topics 25 and 28 we explicitly find the term "agriculture" (*agriculture*), with other keywords related to this field such as "développement"(*development*), "rural"(*rural*). A second interesting subset is the one composed by topics 5, 35, 36 and 39. These topics are represented by keywords about food security and insecurity such as "sécurité"(*security*), "insécurité"(*insecurity*), "alimentaire" (*alimentaire*) highlighting that articles on this issue were discussed in national news. Similarly, topics 26 and 29 seem to be about studies about food security: the word pair "enquête" (*survey*) and "menage"(*household*) was detected, probably referring to the permanent agricultural survey performed annually by the Ministry of Agriculture to compute indicators to prevent food insecurity in the country.

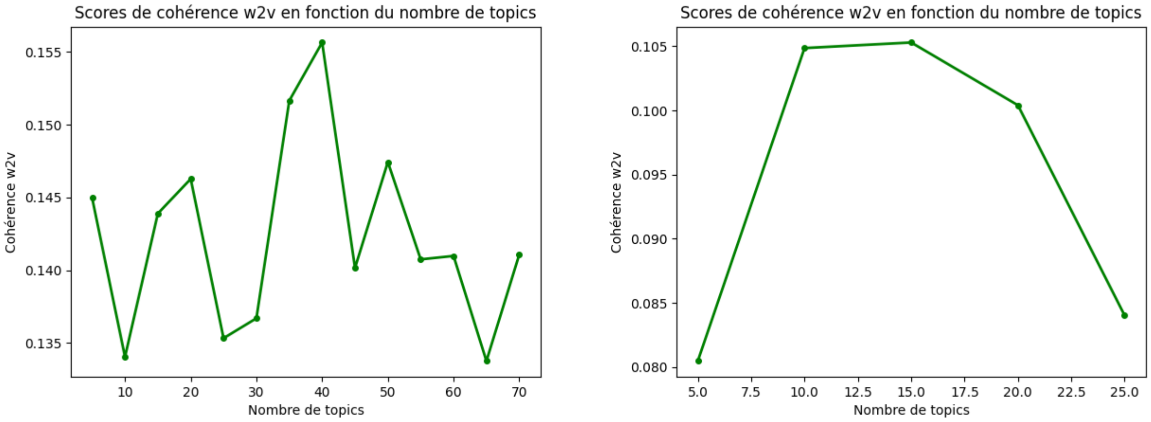
Next, we show the results applied to the YouTube corpus. The best configuration was obtained, as in the previous case, by using TF based vectorization, and filtering terms with minimum DF of 0.05 and maximun of 0.86. Here coherence metrics (Fig. 3b), suggest that the best models are the ones with lower number of topics. However, subsequent verification highlighted that there is a lack of food security related topics in the model. We therefore select the configuration with 25 topics, that shows a higher number of topics related to food security. In Table 2 we present a selection of the topics that we detected as most relevant for food security.

We observe that topic 3 is represented by keywords related to Russia-Ukraine conflict, such as "guerre"(war) and "crise"(crisis), the name of the two nations involved "russie" (Russia) and "ukraine" (Ukraine); it's interesting that also "sénégal" country seems to be cited in the articles about the event. As our methodology detects topics from Food Security articles, this means that the videos may discuss the impacts of the Russia-Ukraine conflict on the food security of the country. So even on this dataset we can observe how topic modeling can help in the detection of events that can explain the signals from warning systems. The other topics selected, 11 and 12, are represented by keywords like *céréale* (cereal) and *blé* (grain) that are more related to the agricultural sector, similarly to what emerged in the topic model for News Articles.

4.2 *TXT-FS* compared to survey based indicators

In this section, we will analyze how the proposed *TXT-FS* indicator (cf. Sec. 3.5) behaves when compared to classic survey based food security indicators. In order to focus on the situation in Burkina Faso, *TXT-FS* on YouTube transcriptions has been computed on a corpus including 283 YouTube transcriptions extracted from two national news channels of Burkina Faso, Burkina24 and RTB - Radiodiffusion Télévision du Burkina.

Figure 4 shows 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. The values of all indicators has been normalized in the range $[0, 1]$, provinces colored in grey correspond to no data cases (e.g., they are not mentioned in the corresponding corpus used to compute *TXT-FS*). Food Consumption Score (*FCS*) and Household Dietary Diversity Score (*HDDS*) are two well known indicators that are computed based on answers to household surveys. These metrics, which are widely used in the scientific literature and by governmental and nongovernmental organizations [14, 26], can be used to assess the frequency, quantity and quality of food in a certain area. The aim of *FCS* is to estimate the cumulative frequency of the different food groups consumed over a period of 7 days within each household taken into account in the survey. *FCS* can then be considered a proxy of the quantity of nutrients and energy intake. *HDDS* measures food consumption frequency and diversity by focusing on the nutritional quality of the diet, and it is calculated based on the number of different food groups consumed in the last 24 hours. We calculated the values of *FCS* and *HDDS* by using data from the permanent agricultural survey conducted by



(a) Coherence of topics from news

(b) Coherence of topics from videos

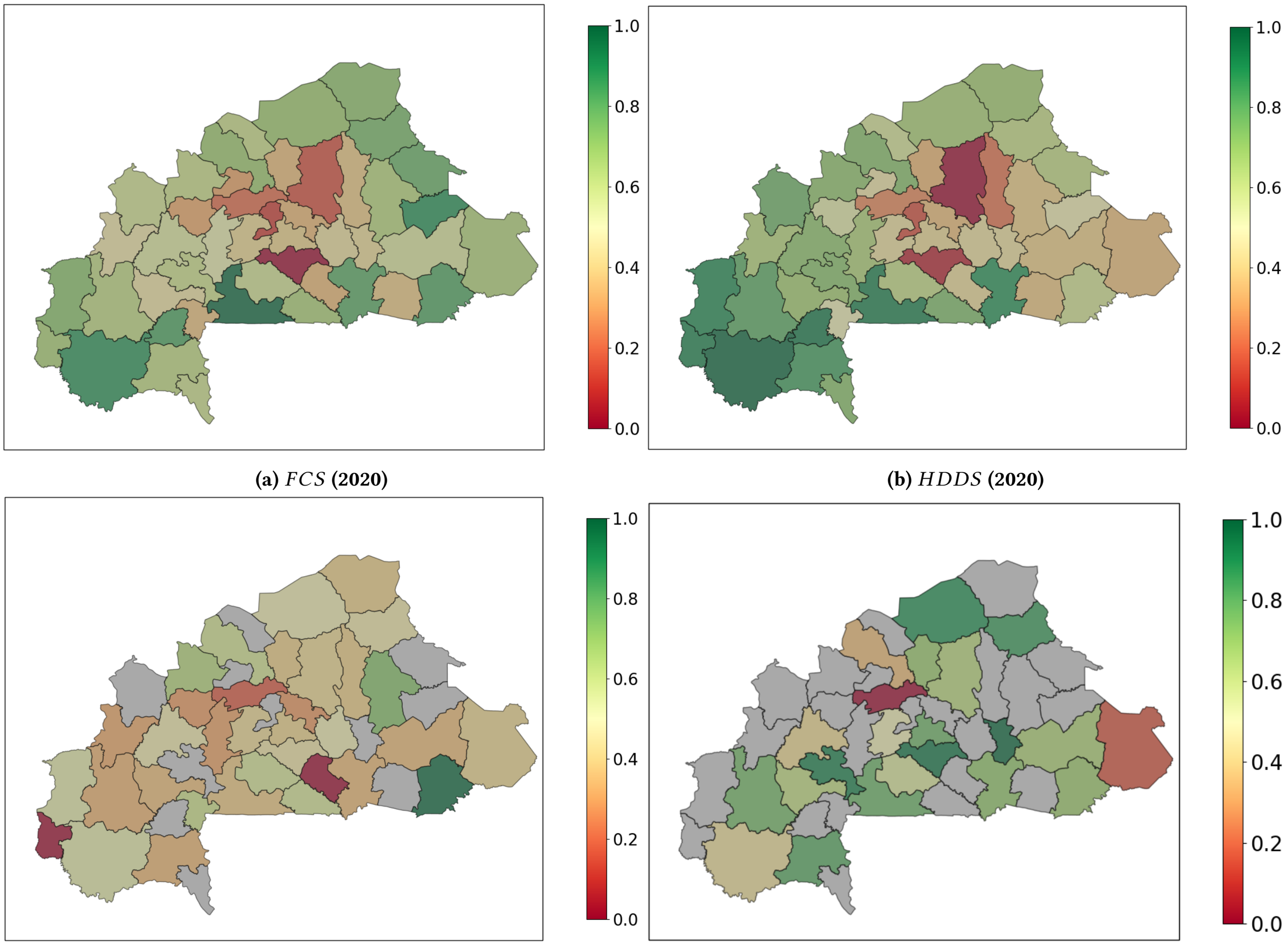
Figure 3: Coherence evaluation for different number of topics given the best configuration. In (a) coherence values using from 5 to 70 topics on the News articles; In (b) coherence values using from 5 to 25 topics on the video dataset.

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

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



(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.

the Burkinabè government, that was kindly provided to us by the Ministry of Agriculture of Burkina Faso, as for the work in [4].

It can be seen from Figure 4 how some patterns regarding the food security of some provinces are consistent among different indicators. For instance, the *Passoré* province (red one in the center-north of the country, in Fig. 4d) shows low food security according to all the four maps. The *Zoundwéogo* province (red one in the center-south of the country, in Fig. 4c) has a relatively critical situation according to both *FCS* and *HDDS*, confirmed by *TXT-FS* on news articles. To name another example, the *Tapoa* province (eastern one on the right) shows low values for *HDDS* and both versions of *TXT-FS*. Several other cases of consistency can be found in the maps. While the general depiction of the food security situation is similar across the four maps,on the other hand, some counter-examples with the same province showing a more or less critical situation according to different indicators can also be found: this can be considered completely normal and in line with the food security analysis context. As already pointed out, such a complex phenomenon needs to be measured from different angles (i.e., by using different indicators), that may give complementary information about a certain area. In fact, it can also be noted how even the two indicators based the same household survey data (*FCS* and *HDDS*) show some discrepancies in measuring the food security of some areas. Note also how a certain temporal bias in introduced by the fact that, while *FCS* and *HDDS* are computed based on household surveys conducted in 2020, *TXT-FS* is based on textual data covering a larger timespan, especially regarding new articles.

5 CONCLUSIONS

In this work, we proposed a pipeline for the spatio-temporal analysis of textual data to support the explanatory process of food insecurity situations. The proposed pipeline integrates different textual analysis approaches to obtain an explanatory model evaluated on real-world and large-scale data. We also proposed a food security indicator based on textual data, namely *TXT-FS*. We focus on the area of West Africa, which has been at the center of many humanitarian crisis since the beginning of this century, with particular attention on the country of Burkina Faso. The results of our analyses have proven how our approach provides significant results that offer distinct and complementary qualitative information on the subject of food security and its spatial and temporal characteristics.

While this work poses some fundamental step towards the use of textual data for food security analysis, the process has still room for improvements. The discovery and selection process of the locations associated to each textual document can be done with a finer grain (e.g., at the sentence level) in order to properly process documents discussing facts happening in several locations. Also, we plan to extend the use of BERT-based language models for different steps of the pipeline, e.g., topic modeling. In the future, we also plan to perform analyses that explicitly take into account the time dimension, in order to study the food security dynamics over several years.

Regarding the data, we plan to extend and refine the lexicon about crisis situations, and we are currently collecting a much larger corpus of textual documents over three West African countries (Burkina Faso, Benin and Senegal).

ACKNOWLEDGMENTS

This work was supported by the French National Centre for Space Studies (CNES), as part of the FRESA (Fouille de texte pour la REcherche de consensus sur la Sécurité Alimentaire) project, APR TOSCA 2021. This study was conducted with the help of the Ministry of Agriculture, Water Resources, Sanitation and Food Security of Burkina Faso, which provided data from the permanent agricultural survey. This work was also supported by the French National Research Agency under the Investments for the Future Program #DigitAg, referred to as ANR-16-CONV-0004. We also want to thank Hugo Deléglise for the preliminary studies about text mining applications for food security.

REFERENCES

[1] D. M. Blei, A. Ng, and M. I. Jordan. 2003. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* 3 (2003), 993–1022.

[2] Burkina Faso Government. 2009. *Inondations du 1er Septembre 2009 au Burkina Faso. Evaluation des dommages, pertes et besoins de construction, de reconstruction et de relèvement.* Burkina Faso Government, FAO.

[3] J. Coates. 2013. Build it back better: Deconstructing food security for improved measurement and action. *Global Food Security* 2 (2013), 188 – 194.

[4] H. Deléglise, R. Interdonato, A. Bégué, E. Maître d'Hôtel, M. Teisseire, and M. Roche. 2022. Food security prediction from heterogeneous data combining machine and deep learning methods. *Expert Systems with Applications* 190 (2022), 116189.

[5] H. Deléglise, M. Roche, R. Interdonato, M. Teisseire, A. Bégué, and E. Maître d'Hôtel. 2022. *Automatic extraction of food security knowledge from newspaper articles - Appendix.* Working Paper. Agritrop.

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

[7] H. Deléglise, C. Schaeffer, E. Maître d'Hôtel, A. Bégué, M. Roche, R. Interdonato, and M. Teisseire. 2021. Corpus de journaux burkinabés en français sur la sécurité alimentaire publiés entre 2009 et 2018. https://doi.org/10.18167/DVN1/IVVEQL

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv* abs/1810.04805 (2019).

[9] B. Drury and M. Roche. 2019. A survey of the applications of text mining for agriculture. *Comput. Electron. Agric.* 163 (2019).

[10] FAO, FIDA, OMS, WFP, and UNICEF. 2020. *The State of Food Security and Nutrition in the World - Transforming Food Systems for Affordable Healthy Diets.* FAO.

[11] P. Foini, M. Tizzoni, D. Paolotti, and E. Omodei. 2021. On the forecastability of food insecurity. *medRxiv* (2021).

[12] J. Hoddinott. 1999. *Choosing Outcome Indicators Of Household Food Security, Vol. Technical Guide No 7.* International Food Policy Research Institute.

[13] R. Interdonato, J.-L. Guillaume, and A. Doucet. 2019. A lightweight and multilingual framework for crisis information extraction from Twitter data. *Soc. Netw. Anal. Min.* 9, 1 (2019), 65:1–65:20.

[14] A. D. Jones, F. M. Nguren, G. Pelto, and S. L. Young. 2013. What Are We Assessing When We Measure Food Security? A Compendium and Review of Current Metrics. *Advances in Nutrition* 4 (2013), 481–505.

[15] D. Kasprzyk. 2001. *Measurement error in household surveys : sources and measurement.* United States Federal Committee on Statistical Methodology.

[16] Quoc V. Le and T. Mikolov. 2014. Distributed Representations of Sentences and Documents. In *ICML*.

[17] J. M Lepkowski. 2001. *Non-observation error in household surveys in developing countries.* United States Federal Committee on Statistical Methodology.

[18] J. Leskovec, A. Rajaraman, and Jeffrey David Ullman. 2020. *Mining of massive data sets.* Cambridge university press.

[19] C. Manning, P. Raghavan, and H. Schütze. 2010. Introduction to information retrieval. *Natural Language Engineering* 16, 1 (2010), 100–103.

[20] L. Martin, B. Muller, P. J. Ortiz Suárez, Y. Dupont, L. Romary, É. de la Clergerie, D. Seddah, and B. Sagot. 2020. CamemBERT: a Tasty French Language Model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.* 7203–7219.

[21] G. Martini, A. Bracci, S. Jaiswal, M. Corea, L. Riches, J. Rivers, and E. Omodei. 2021. Nowcasting food insecurity on a global scale. *medRxiv* (2021).

[22] B. D. Meyer, W. K.C. Mok, and J. X. Sullivan. 2015. Household surveys in crisis. *Journal of Economic Perspectives* (2015).

[23] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *NIPS*.

[24] OCHA. 2010. *Burkina Faso ; Inondations ; période du 1er au 30 septembre 2010.* United Nations Office for the Coordination of Humanitarian Affairs.

[25] M. Röder, A. Both, and A. Hinneburg. 2015. Exploring the Space of Topic Coherence Measures. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining* (2015).

[26] E. Vhurumuku. 2014. *Food security indicators - FAO.* Integrating Nutrition and Food Security Programming for Emergency response workshop. Technical report.