Mining news articles dealing with food security

Hugo Deléglise, Agnès Bégué, Roberto Interdonato, Elodie Maître D'hôtel, Mathieu Roche, Maguelonne Teisseire

Mining news articles dealing with Food Security

Hugo Deléglise[1,3], Agnès Bégué[1,3], Roberto Interdonato[1,3], Elodie Maître d'Hôtel[2,3], Mathieu Roche[1,3], and Maguelonne Teisseire[1,5]

[1] TETIS, Univ Montpellier, AgroParisTech, CIRAD, CNRS, INRAE, Montpellier, France.
[2] MOISA, Univ Montpellier, CIHEAM-IAMM, CIRAD, INRAE, Institut Agro, Montpellier, France.
[3] CIRAD, UMR TETIS, F-34398 Montpellier, France.
[4] CIRAD, UMR MOISA, F-34398 Montpellier, France.
[5] INRAE, Montpellier, France.

Abstract. Food security is a major concern in West Africa, particularly in Burkina Faso, which has been the epicenter of a humanitarian crisis since the beginning of this century. Early warning systems for food insecurity and famines rely mainly on numerical data for their analyses, whereas textual data, which are more complex to process, are rarely used. To this end, we propose an original and dedicated pipeline that combines different textual analysis approaches (e.g., word embedding, sentiment analysis, and discrimination calculation) to obtain an explanatory model evaluated on real-world and large-scale data. The results of our analyses have proven how our approach provides significant results that offer distinct and complementary qualitative information on the food security theme and its spatial and temporal characteristics.

Keywords: Food security · sentiment analysis · spatiotemporal analysis · term discrimination · text mining · word embedding

1 Introduction

Hunger remains a major problem in many parts of the world. Although a large scale and permanent solution to this situation is far from being achieved, steady progress was made in the first 15 years of this century. Among West African countries, Burkina Faso is in one of the most severe situations, with an undernourishment prevalence of 21.3 % from 2015-2017 [7]. Burkina Faso is also one of the countries most affected by the phenomenon commonly known as the "triple burden of malnutrition", characterized by the coexistence of overnutrition, undernutrition and micronutrient deficiencies in the population.

Following several food crises in the 1970s and 1980s in different regions of the world, several food security alert and monitoring systems (FSMSs) were created by governmental organizations and NGOs. The objective of these systems, which are still very active today, is to prevent food crises and to help countries plan food aid programs to optimize their food production and distribution channels. In this study, we examine the ability of text mining methods to extract and

analyze the qualitative information used as proxies for the national and regional food situation and its evolution over the last ten years in Burkina Faso from a corpus of newspapers from the country. The aim is to provide explainable indicators complementary to the automatic predictions of food security scores, i.e., as the ones obtained in our previous works based on the application of machine learning approaches on heterogeneous [3] and textual [2] data.

However, the difficulties associated with the implementation of text-mining approaches are linked to the structural complexity of textual data and are the subject of a large number of studies. We now detail the approaches proposed in the scientific literature to address these difficulties. In the domain of agriculture, which is closely related to food security, information extraction from textual data is a topic that has been attracting increasing interest [6]. In this field, several studies have focused on sentiment analysis [16], named entity extraction (i.e., places, dates or individuals related to agriculture) [12], etc.

The originality and methodological contributions of this work are presented at 3 levels: (1) its multidisciplinary aspect involving the combination of approaches based on text mining (e.g., word embedding and sentiment analysis) for the analysis of food security, which has been little studied from this perspective; (2) spatiotemporal analysis based on the content of French texts; and (3) extension of discrimination measures to address spatiotemporal data. The usefulness of this approach is to propose an explanatory framework complementary to the outputs of the predictive models usually applied to other types of data (e.g., digital data and satellite images). While we focus on the study case of Burkina Faso, the proposed method is generic and can be applied to any other area in the world. Section 2 presents the proposed approach. Section 3 outlines and discusses the information extracted from a dedicated corpus in French.

2 Proposed approach

2.1 Text mining approaches

Studies use text mining methods to extract information on food security-related events from newspapers which proposes a framework for automatic detection of food crises [19]. Their method consists of extracting the most characteristic vocabulary (keywords) by tf-idf (term frequency-inverse document frequency) for each article [15], which is a method of weighting characteristic terms of texts, and then extracting the named entities with a Bi-LSTM-CNN-CRF framework [20]. A weight is associated with each keyword according to its semantic similarity (by Word2vec) with the terms of the article title. Each article, through a set of weighted keywords and associated named entities, is classified by single-pass clustering [14]. The ability of tf-idf to extract relevant and specific vocabulary from newspaper articles has also been demonstrated [1]. In this context, some text-mining methods are integrated in our pipeline:

Word2vec. Word2vec (w2v) [13] is a family of automatic language processing models for word embedding, i.e., the transformation of terms and texts into

vectors. W2v is based on two-layer neural networks and aims at learning vector representations of terms in texts so that terms that share similar contexts (i.e., are often surrounded by the same terms) are represented by close numerical vectors. In our study, a CBOW (continuous bag of words) architecture is used (preferred to the skip-gram architecture, which requires more execution time while sometimes offering less satisfactory performance for processing newspaper articles [10]). CBOW aims at predicting the appearance of a term by using as proxies the terms that are close to it in the text. The model is trained on a large training corpus (French Wikipedia, in our study) by traversing each term and its neighbors and obtaining a set of feature vectors that represent each term in the text as the output.

Term polarity. The polarity of a term is a criterion that indicates whether it is positive, negative or neutral [17,9]. In our context, the average polarity of texts dealing with food security can give us relevant information about their worrisome or even alarming character. There are currently few methods for performing sentiment analysis on French texts. To evaluate the negativity of a term, we use the French version of the sentiment analysis model VADER (Valence Aware Dictionary and Sentiment Reasoner) implemented by the Python package vaderSentiment-fr[6]. This model is based on a lexicon of 7500 terms classified as positive or negative and on contextual rules that can modify the valence of the terms (e.g., the use of negation, punctuation, capitalization, and adverbs). This model was chosen because it has a good compromise between its simplicity of implementation and execution time and its classification performance, performing better than many existing methods, some of which are based on the use of machine learning [8].

tf-idf. To evaluate the discrimination of the terms of an article, we use the concept of tf-idf (term frequency-inverse document frequency) [15], which measures to what extent a term is characteristic of a text by evaluating its relevance and its singularity. Its principle is based on a formula in which two values, tf (term frequency) and idf (inverse document frequency), are multiplied together. tf corresponds to the frequency of a term in a text, and it therefore increases when a term is frequent in the text. idf measures the importance of a term according to its distribution in all the texts studied rather than based on its frequency in a particular text.

2.2 Our food security pipeline

The objective of our pipeline is to perform a spatiotemporal analysis of food security based on the terminology of this domain linked to the textual proxies we define. In this framework, we propose an original and dedicated pipeline that combines different textual analysis approaches (e.g., word embedding, sentiment analysis, and discrimination calculation). To this end, we present in this section the methodology deployed to obtain a spatial and temporal explanatory context of the Burkinabe food situation from the corpus of newspapers studied. Fig. 1

[6] https://pypi.org/project/vaderSentiment-fr/

summarizes the analysis plan. For this a first general lexicon on food security is therefore used to detect articles of interest, we name this lexicon "$GLEX$" (Generalist LEXicon). Then, two other more detailed lexicons are used to detect the expressions of "food security" and "crisis" themes used and thus obtain a more qualitative view of the content of the articles. We call these two detailed lexicons on food security and on crises "$FLEX$" (Food LEXicon) and "$CLEX$" (Crises LEXicon), respectively. These lexicons are freely available [5].

First, we present step **(1)** of selecting relevant articles. For this, we compute by w2v the semantic similarity between each article and the generalist lexicon $GLEX$, used as a basis to identify articles on the theme "food security". The principle is to consider an article as dealing with food security if its semantic similarity with $GLEX$ by w2v is higher than a threshold x (chosen and validated in the Appendix document [4]). This aims to detect the articles of interest to focus the analyses.

Second, we establish in step **(2)** the textual proxies of food security on the selected articles. To this end, we perform the following operations:

- We keep for the selected articles their w2v score calculated during step **(1)**, which quantifies their degree of connection with the food security theme and constitutes a proxy of this domain.
- We compute the negativity rate of the articles we propose as a proxy, i.e., the frequency of the negative terms in each article (Formula 1), to obtain information on the alarming nature of the articles' content.

$$Neg(art) = \frac{nb_{terms\_neg}(art)}{nb_{terms}(art)} \tag{1}$$

  where *Neg* is the negativity rate of an *art* article, and nb_{terms_neg} and nb_{terms} represent the number of negative terms and the number of terms of an *art* article, respectively, based on the French version of the VADER model (Valence Aware Dictionary and Sentiment Reasoner).
  The hypothesis assumes that articles published during periods and in areas of food insecurity are associated with more negative valences than in a context of food sufficiency. An article is considered to be negative if its negativity rate is greater than 0.1 (the threshold validation methodology is detailed in the Appendix document [4]).
- We study the most used vocabulary in articles related to food security to detect whether the vocabulary adopted is consistent with the trends and crises that have affected food security in the country and thus to have a more explanatory perspective of the data. To accomplish this, we calculate for each article the frequency of 119 expressions from the two detailed lexicons $FLEX$ and $CLEX$.

Third, we describe step **(3)** of global, regional and annual analysis of the proxies defined in step **(2)**. To take into account the spatiotemporal aspect of food security, the proxies presented are then aggregated at different granularities to perform targeted analyses at the global, regional and annual levels and thus

be able to visualize the trends and food crises that have affected the country over the last decade. The proxies are aggregated at three levels:

Global level: this level of analysis provides a general view of the characteristics of the country's food situation between 2009 and 2018 and can be used as a comparison for targeted analyses (regional and annual). The proportions of articles dealing with food security and negative articles are calculated over the entire corpus. We consider the average frequency of occurrence of each term in the detailed $FLEX$ and $CLEX$ lexicons across all articles in the corpus.

Regional level: this level of analysis aims to provide a representation of the food situation and its characteristics at the regional level. We illustrate our analyses with three regions: the Centre, Hauts-Bassins and Sahel regions. These three regions were chosen because they are among the most frequently cited in the articles in the corpus and are associated with distinct health situations [18]. Our approach consists of considering an article as associated with a region if a locality of the region is mentioned at the beginning of the article (i.e., in the title or in the first sentence of the article). The proportions of articles dealing with food security and negative articles were calculated for each of the 3 regions. To extract the characteristic regional vocabulary, we compute the tf-idf of each term of the $FLEX$ and $CLEX$ lexicons on the articles of each considered region. In our context, tf-idf allows us to highlight the expressions of food security and crises that are frequent in the articles related to a certain region and that are more specifically used in the articles of the region (i.e., more than for the other articles).

Annual level: this level of analysis provides annual characteristics of the food situation in Burkina Faso and tracks its evolution from 2009 to 2018. Each article is associated with its year of publication, which is extracted in the metadata linked to the article. This proposal, called the TIR (Tf-Idf ratio), is based on the concept of tf-idf and proves to be more suitable in our context, allowing us to distinguish rare and year-specific expressions more than tf-idf. More precisely, we first compute for each expression of the lexicons $FLEX$ and $CLEX$ the tf-idf of the expression on average on the articles of the year (Formula 2); then, in a second step, we compute the ratio of this tf-idf by the tf-idf of the expression on average on the articles of other years (TIR ratio, (Formula 3)).

$$TF - IDF_{moy}(t, A_y) = \frac{\sum_{art \in A_y} TF - IDF(t, art)}{N_y} \tag{2}$$

where $TF-IDF_{moy}$ is the average tf-idf of the term "t" on the articles "art" belonging to the set A_y of the articles of year y; we note N_y is the cardinality of this set.

$$TIR(t, A_y) = \frac{TF - IDF_{moy}(t, A_y)}{TF - IDF_{moy}(t, A_z)} \tag{3}$$

where TIR is the ratio of the tf-idf of the term "t" averaged over the articles belonging to the set A_y of articles in year y to the tf-idf of the term "t" averaged over the articles belonging to the set A_z of articles in different years of year y.

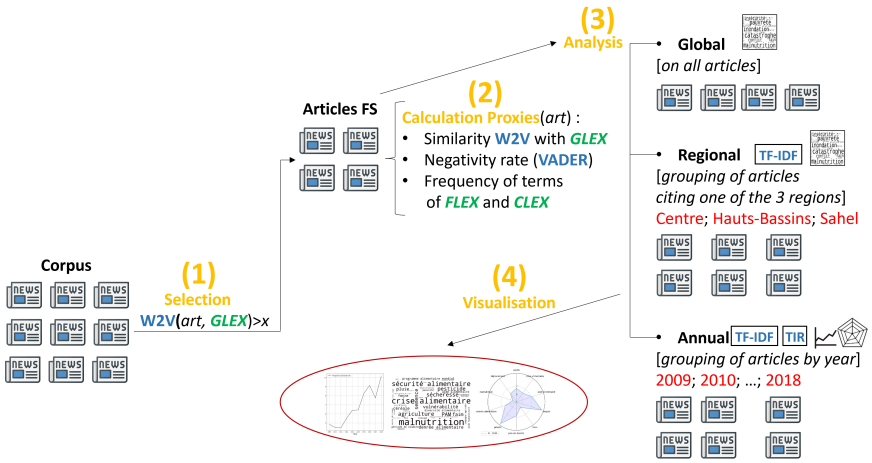


Fig. 1. General illustration of the analysis plan. The main steps are numbered: (1) selection of articles of interest; (2) computation of textual proxies on these articles; (3) global, spatial and temporal analysis; and (4) visualization.

3 Experiments

3.1 Corpus of newspapers

Currently, the main Burkinabe newspapers have their own news website on which they publish their articles. For the creation of our newspaper corpus, we turned to two Burkinabe newspapers whose websites allow for good data accessibility: Burkina24 and LeFaso. These newspapers are among the most read newspapers in the country and have a large number of articles on various topics online. We extracted a total of 22856 articles between 2009 and 2018 (5595 for Burkina24 and 17261 for LeFaso), a period during which food security has undergone significant variations and several crises. The articles were filtered and then lemmatized with the Python package *Spacy* [7].

3.2 Results

Regional Analysis. We focus here on certain regions and observe whether the food security proxies aggregated over these regions are associated with the known regional food situation. The three regions studied are the Centre, Hauts-Bassins and Sahel regions.
In Table 1, we see that the Hauts-Bassins region, which is the least food insecure of the regions presented, is associated with the lowest proportions of "food security" theme articles and negative articles. Conversely, the Sahel region, which is

[7] https://spacy.io/api/lemmatizer

experiencing the most severe situation, has the highest proportions of "food security" themed articles and negative articles, which are both significantly higher than those at the national level. These data are consistent with expectations: the more an area is plagued by food insecurity and/or crises, the more articles will mention these topics and be negative.

In Fig. 2, we see that for the Hauts-Bassins region, the least poor, the most important expressions of the lexical field of food security are neutral (e.g., "riz" (*rice*), "agriculture" (*agriculture*), and "campagne agricole" (*agricultural campaign*)), whereas in the Sahel region, the expressions of food security are more negative (e.g., "malnutrition" (*malnutrition*) and "crise alimentaire" (*food crisis*)). Regarding the expressions of the theme of crises, the word clouds highlight concerns that are characteristic of each region. For example, the expression "inondation" (*flood*) is the most important in the Centre region, which is plagued by this problem, whereas the expression "foncier" (*land*) is the most important in the Hauts-Bassins region, where land management is a major problem.

Table 1. Comparison of the percentage of articles on the theme of "food security" (FS) and the percentage of negative articles for the three regions of the Centre, Hauts-Bassins, and Sahel and for Burkina Faso (BF).

	Centre	Hauts-Bassins	Sahel	BF
% of FS theme articles	6.5	4.9	10.7	7.3
% of negative articles	4.8	1.3	12.1	6.4

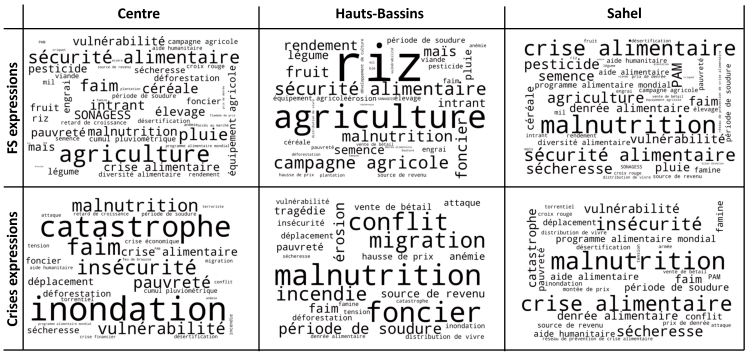


Fig. 2. Word clouds of expressions from the *FLEX* (FS expressions) and *CLEX* (Crises expressions) lexicons, based on "food security" theme articles related to three regions (Centre, Hauts-Bassins and Sahel). The size of the terms is proportional to their average tf-idf.

Annual analysis. In this section, we examine whether food security proxies point over time to elements that are consistent, nuanced, or even contradictory to observations and events that have taken place over the past decade that may

have affected food security. Namely, there has been a decline in food security since 2013 as well as events negatively impacting food security (e.g., flooding, drought, and conflict). The objective here is to determine the annual characteristics and evolution of the Burkinabe food situation through the food security proxies considered. We compute for each expression in the $FLEX$ and $CLEX$ lexicons the tf-idf of the expression averaged over the articles of the year, as well as the ratio TIR that we proposed.
Finally we analyze the evolution of the food security and crisis vocabularies used in the articles as a function of time (see Fig. 3a). In Fig. 4 (a), which represents the evolution of the proportion of negative articles by year from 2009 to 2018, we see a trend for negative articles to decrease in proportion. This may seem counterintuitive, and it may be explained by a certain freedom of the press that has tended to decline over the last decade (see Fig. 4 (b)). Fig. 3b shows the evolution of the tf-idf of 5 expressions on average on the articles of each year between 2009 and 2018. We can see an upward trend in the tf-idf of the terms "sécurité alimentaire" (*food security*) and "malnutrition" (*malnutrition*), which have been increasingly used over the last decade. Moreover, some peaks correspond to the year of occurrence of events that took place over the period: the tf-idf of the expression "sécheresse" (*drought*) was the highest in 2012, which experienced a severe drought. The tf-idf for "conflits" (*conflict*) and "déplacement" (*displacement*) peaks in 2013, when conflicts in the Sahel led to the displacement of people from Sahelian countries bordering Burkina Faso.

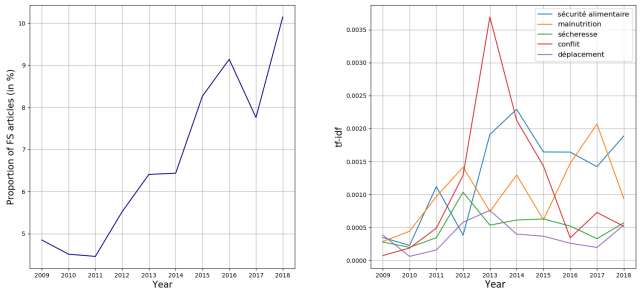
4 Conclusion and Future work

In this study, we examined the ability of text mining methods to extract spatial and temporal thematic information on food security from newspaper articles by examining the context of Burkina Faso.

We proposed, combined and extended, with adapted text mining methods (the Word2vec lexical embedding model, the VADER sentiment analysis model and the tf-idf term importance weighting method) three types of proxies defined on a set of articles, allowing us to obtain distinct and complementary information on the food security theme. This type of approach and the associated results can be exploited as complementary information to the outputs of predictive models (i.e., based on machine and deep learning). Indeed, machine and deep learning models applied to other types of data (e.g., digital data and satellite images) have strong predictive power but often lack explicability and interpretability. These models can then be validated, nuanced or explained by qualitative information from textual data that could make sense to domain experts and advance their understanding of complex food security phenomena.

To improve the thematic search in a finer way than with the word embedding applied in this work with w2v, technologies based on BERT (bidirectional encoder representations from transformers) and trained models for French, such as CamemBERT or FlauBERT, could also be integrated [11].

Acknowledgments. This work was supported by the French National Research Agency under the Investments for the Future Program #DigitAg, referred to as ANR-16-CONV-0004.



(a) (b)

Fig. 3. (a) Change in the proportion (in percentage) of "food security" (FS) theme articles from 2009 to 2018 on the corpus studied. (b) Evolution of the average tf-idf of 5 expressions from the two detailed lexicons $FLEX$ and $CLEX$ between 2009 and 2018.

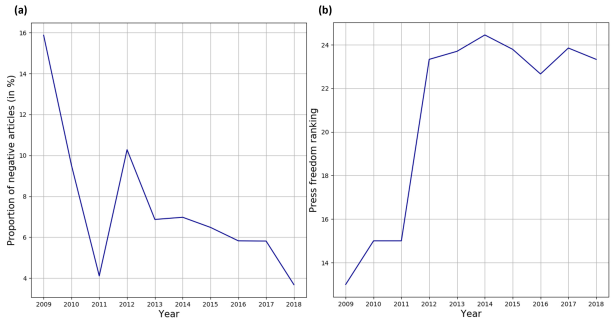


Fig. 4. Changes from 2009 to 2018 in the proportion (in percent) of negative articles per year among articles on the theme of "food security" (a) and Burkina Faso in the press freedom index (Reporters Without Borders[9]) (b).

References

1. Ao, X., Yu, X., Liu, D., Tian, H.: News keywords extraction algorithm based on textrank and classified tf-idf. In: 2020 International Wireless Communications and Mobile Computing (IWCMC). pp. 1364–1369 (2020)

[9] https://rsf.org/fr/methodologie-detaillee-du-classement-mondial-de-la-liberte-de-la-presse

2. Ba, C.T., Choquet, C., Interdonato, R., Roche, M.: 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 (2022)
3. Deléglise, H., Interdonato, R., Bégué, A., Maître d'Hôtel, E., Teisseire, M., Roche, M.: Food security prediction from heterogeneous data combining machine and deep learning methods. Expert Systems with Applications **190**, 116189 (2022). https://doi.org/https://doi.org/10.1016/j.eswa.2021.116189
4. Deléglise, H., Roche, M., Interdonato, R., Teisseire, M., Bégué, A., Maître d'Hôtel, E.: Automatic extraction of food security knowledge from newspaper articles - Appendix. Working paper, Agritrop (2022), `https://agritrop.cirad.fr/600423/`
5. Deléglise, H., Schaeffer, C., Maître d'Hôtel, E., Bégué, A.: Lexiques en français sur la sécurité alimentaire et les crises (2021), `https://doi.org/10.18167/DVN1/C5PU01`, dataverse CIRAD
6. Drury, B., Roche, M.: A survey of the applications of text mining for agriculture. Computers and Electronics in Agriculture **163**, 104864 (2019). https://doi.org/https://doi.org/10.1016/j.compag.2019.104864
7. FAO, ECA: Addressing the threat from climate variability and extremes for food security and nutrition. FAO (2018)
8. Gilbert, C.H.E.: Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: Eighth International Conference on Weblogs and Social Media (ICWSM-14). (2014)
9. Hemmatian, F., Sohrabi, M.K.: A survey on classification techniques for opinion mining and sentiment analysis. Artificial Intelligence Review **52**(3), 1495–1545 (2019)
10. Jang, B., Kim, I., Kim, J.W.: Word2vec convolutional neural networks for classification of news articles and tweets. PloS one **14**(8), e0220976 (2019)
11. Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L., Schwab, D.: Flaubert: Unsupervised language model pretraining for french (2020)
12. Malarkodi, C., Lex, E., Sobha, L.: Named entity recognition for the agricultural domain. In: 17th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2016); Research in Computing Science (2016)
13. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. Proceedings of Workshop at ICLR **2013** (01 2013)
14. Papka, R., Allan, J., et al.: On-line new event detection using single pass clustering. University of Massachusetts, Amherst **10**(290941.290954) (1998)
15. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information Processing & Management **24**(5), 513–523 (1988)
16. Surjandari, I., Naffisah, M., Prawiradinata, M.: Text mining of twitter data for public sentiment analysis of staple foods price changes. Journal of Industrial and Intelligent Information **3** (01 2014). https://doi.org/10.12720/jiii.3.3.253-257
17. Szabolcsi, A.: Positive polarity - negative polarity. Natural Language and Linguistic Theory **22**(2), 409–452 (May 2004)
18. WFP: Burkina Faso : Analyse Globale de la Vulnérabilité, de la Sécurité Alimentaire et de la Nutrition. WFP (2014)
19. Xiao, K., Wang, C., Zhang, Q., Qian, Z.: Food safety event detection based on multi-feature fusion. Symmetry **11**(10) (2019). https://doi.org/10.3390/sym11101222
20. Yu, H.: Named Entity Recognition with Deep Learning. Ph.D. thesis, Auckland University of Technology (2019)