Automatic extraction of food security knowledge from newspaper articles - Appendix*

Hugo Deléglise[a,c], Mathieu Roche[a,c], Roberto Interdonato[a,c], Maguelonne Teisseire[a,e], Agnès Bégué[a,c], Élodie Maître d'Hôtel[b,d]

[a] *TETIS, Univ Montpellier, AgroParisTech, CIRAD, CNRS, INRAE, Montpellier, France.*
[b] *MOISA, Univ Montpellier, CIHEAM-IAMM, CIRAD, INRAE, Institut Agro, Montpellier, France.*
[c] *CIRAD, UMR TETIS, F-34398 Montpellier, France.*
[d] *CIRAD, UMR MOISA, F-34398 Montpellier, France.*
[e] *INRAE, Montpellier, France.*

Abstract

This document is an Appendix of a paper under submission (Deléglise et al., Automatic extraction of food security knowledge from newspaper articles).

1. Validation of the x threshold at which an article is considered to be dealing with food security

We present here the approach adopted to evaluate the potential of w2v to detect articles of interest (i.e., of "food security" theme) and then to set an optimal separation threshold between articles dealing with food security and the remainder of the corpus.

1.1. Evaluation of the relevance of w2v

First, we want to evaluate the ability of w2v to detect articles that deal with food security by testing whether the articles with the highest w2v similarities are the most related to the topic of food security and whether the articles' relationship to food security decreases when their w2v similarity decreases. For this purpose, the articles are sorted by decreasing w2v similarity, and we select a sample of these articles: articles 1, 2, 3, 4, 5, 100, 200, 300, 400, 500, 1000,

*Automatic extraction of food security knowledge from newspaper articles - Appendix

Email addresses: hugo.deleglise@teledetection.fr (Hugo Deléglise), mathieu.roche@cirad.fr (Mathieu Roche), roberto.interdonato@cirad.fr (Roberto Interdonato), maguelonne.teisseire@inrae.fr (Maguelonne Teisseire), agnes.begue@cirad.fr (Agnès Bégué), elodie.maitredhotel@cirad.fr (Élodie Maître d'Hôtel)

1500, 2000, 2500, 3000, 4000, 5000, 6000, 7000, and 8000 to obtain a panel of 20 articles made up of both articles with the highest w2v scores and articles with low w2v scores, which are not supposed to be related to food security. The content of these 20 articles is available on GitHub[1].

Each article in this panel is then manually annotated according to whether it addresses food security to serve as a comparison to the scores assigned by w2v and to be able to assess the quality of these scores. The agreement between the annotators' choices is therefore crucial and is also evaluated in this section. An annotation guide (whose methodology is detailed below; available as a pdf version on GitHub[1]) was designed to allow stakeholders to annotate in a standardized, consistent manner and in agreement with each other.

A preliminary stage is first devoted to creating the criteria for the annotation classes and to tuning the annotators with these criteria by group study of a small sample of articles.

Then, the 20 articles from the selected panel were anonymized (i.e., separated from their associated w2v score) and manually annotated by three food security experts. The classes to be annotated are the following:

- "0" : the article does not deal with food security, nor with any related theme;
- "1" : the article deals with a theme related to food security, i.e., related to events that can indirectly improve or worsen the food situation in the population (e.g., climate, locust swarm, poverty, massive layoff of workers, and new methods of agriculture);
- "2": the article directly addresses one of the 4 pillars of food security (availability of food; access of populations to food; event that directly disrupts the stability of access to food (with mention in the article of the food consequences of the event); good or bad use of food at the health and nutritional level).

To ensure a greater robustness of the annotations, the majority class is chosen. More precisely, for each article, the most chosen class among the three annotators is assigned; in the case of an article for which each annotator chooses a different class (i.e., "0", "1", and "2"), it is the median, i.e., class "1", that is assigned.

In Table 1, we see as expected a clear tendency for articles with the highest w2v similarities to be classified "1" or "2" and articles with lower w2v similarities

[1] https://github.com/pipapou/20_articles_BF

to be classified "0". Among the 5 articles with the highest w2v scores in the panel, 4 have a majority class of "2". Of the 10 articles with the highest w2v scores in the panel, 8 have a majority class of at least "1". Conversely, 9 of the 10 articles with the lowest w2v scores had a majority class of "0". We deduce that the use of w2v to detect and classify articles dealing with food security is relevant in our context. Moreover, we can note a significant agreement between annotators' choices: 13 articles out of 20 (i.e., 65%) were identically labeled by the 3 annotators, and the 20 articles (i.e., 100%) were identically labeled by 2 annotators among 3. Finally, the Fleiss Kappa [1] is computed, and this coefficient is a statistical measure of the agreement of labeling between several annotators (" < 0" if the agreement is nonexistent, " > 0" if the agreement is at least weak and " $= 1$" if the agreement is perfect). The Fleiss Kappa in our case is equal to 0.601, which means a significant agreement between the annotators' choices. Despite the small sample size, the p value associated with the Kappa is less than 0.01, which allows us to conclude that it is significant. We conclude that the manual annotation performed by the 3 annotators is relevant to serve as a reference for the scores assigned by w2v and allows us to appreciate the quality of these scores.

Table 1: Comparison of classes assigned by three experts to 20 articles in the corpus and majority classes.

Rank w2v	**Expert classes 1, 2 & 3**	**Majority class**
1	2 | 2 | 1	2
2	2 | 2 | 2	2
3	2 | 2 | 2	2
4	0 | 0 | 0	0
5	2 | 2 | 2	2
100	1 | 0 | 1	1
200	0 | 1 | 1	1
300	2 | 1 | 1	1
400	1 | 0 | 1	1
500	0 | 0 | 0	0
1000	0 | 0 | 0	0
1500	0 | 0 | 0	0
2000	0 | 0 | 0	0
2500	0 | 0 | 1	0
3000	0 | 1 | 0	0
4000	0 | 0 | 0	0
5000	0 | 0 | 0	0
6000	0 | 0 | 0	0
7000	0 | 0 | 0	0
8000	1 | 1 | 1	1

1.2. Identification of potential w2v thresholds

In a second step, we want to define a threshold w2v x such that articles with similarity w2v higher than x (resp. lower than x) are automatically classified as dealing with food security (resp. not dealing with food security). If this threshold is too low, a significant proportion of the selected articles will not be relevant. If the threshold is too high, too few articles will be selected for relevant analysis. A balanced approach must be found on the level of constraints to be set to consider an article as dealing with food security.

To take this into account, we first construct an interval within which the final w2v threshold will be chosen. The lower bound is set empirically. By checking (by reading) for several "threshold" values of approximately 30 articles whose w2v score is just above the value, the value is chosen as the lower bound if at least a quarter of the checked articles are related to food security. We start with the value 0.1 and then advance in steps of 0.05. The first value tested for which we have at least a quarter of food security-related articles is 0.3. To obtain as large an interval of potential thresholds as possible, we also tested the threshold 0.28, located between 0.3 and 0.25, which had been tested as insufficient. For this threshold of 0.28, more than a quarter of the articles are considered relevant, so we validate this value as a lower bound. For the selection of the upper bound, whose associated threshold should allow for the selection of enough articles for the analyses, we must carefully examine the number of articles selected by the choice of threshold that are associated with each region studied (i.e., Centre, Hauts-Bassins, Sahel). Indeed, contrary to the years that can be associated with each article (because the information on the year of publication is present in the metadata), a minority of articles in the corpus are associated with at least one of the 3 regions studied (25 %). We consider an upper bound as acceptable as long as it allows us to associate at least 50 articles with each region. This number of 50 is determined empirically, and there is no minimum sample selection rule for the analyses (some of which are complex) that we will apply (e.g., detection of specific vocabularies in the articles and calculation of their tf-idf values and TIR ratios). We check for several values that this minimum size is verified for each region, starting with 0.28 (the lower bound) and advancing by steps of 0.02. Table 2 details the number of articles selected as dealing with food security across the entire corpus and for each of the 3 regions studied as a function of the value chosen as the detection threshold (0.28, 0.30, 0.32, 0.34, 0.36 and 0.38). The last value tested for which we have numbers related to each region of size at least 50 is 0.36, so we choose this value as the upper bound.

Table 2: Illustration of the number of articles selected as dealing with food security in the entire corpus and for each of the three regions studied (Centre, Hauts-Bassins, Sahel) according to the value chosen as the detection threshold. Values in red correspond to numbers below 50.

Threshold	**Corpus**	**Centre**	**Hauts-Bassins**	**Sahel**
0.28	6174	1098	320	192
0.30	4694	813	242	153
0.32	3451	574	124	166
0.34	2472	389	114	93
0.36	1675	252	74	66
0.38	1068	152	46	44

1.3. Choice of the w2v threshold

To validate the threshold x, we propose to set 3 thresholds in the interval [0.28,0.36] that we have constructed and to check whether the articles corresponding to the critical areas of these thresholds (articles with w2v similarity just below or above the threshold) are relevant for food security. For this purpose, we set thresholds x to be evaluated: $x = 0.28$; 0.32; 0.36, homogeneously fixed between the two "limit" values of the interval.

For each of the three x thresholds, we select the 12 articles with w2v similarity greater than x that are closest (considered to be related to food security) and the 12 articles with w2v similarity less than x that are closest (considered not to be dealing with food security). This gives a total of 24 articles, of which 50% are above the x threshold and are therefore considered to deal with food security. Note that this number of 24 articles evaluated by threshold is insufficient to compare the quality of the thresholds in a statistically significant way. We carried out a χ^2 test of equality of the proportions of errors associated with the 3 thresholds, which proved not to be significant (p value > 0.05). This annotation work needs to be extended (by at least a hundred articles per threshold) to obtain statistically representative and comparable samples. This additional investment on the part of the experts involved will be judicious to consolidate the choice of threshold detailed below. The statistics presented hereafter to validate the x threshold therefore retain trends and should be confirmed by other more thorough studies.

The 24 selected articles (for each of the three x thresholds) were then manually annotated with the same methodologies, annotation guides and classes as presented above.

We choose the x threshold for which the 24 associated articles were most frequently concordantly classified by w2v thresholding against the manual annotation. We concordantly classified those articles with lower w2v similarity

(resp. higher) than x and manually annotated as "0" (resp. "1" or "2") and consider the corresponding F-measure as a criterion for choosing x.

Table 3 represents for each threshold x the percentage distribution of manually annotated classes ("0", "1", and "2") over the 24 corresponding articles, as well as the error, recall, precision, and F-measure rates (defined in (1)) of the w2v thresholding with respect to manually annotated classes. We observe that increasing the x threshold does not significantly decrease the percentage of food security irrelevant articles (manually annotated as "0") at the threshold contours but does increase the percentage of highly relevant articles (classified as "2"), from 0% (for x=0.28) to 12.5% (for x=0.36). Finally, we note that the F-measure is maximized for x=0.36, so we choose this threshold. This threshold w2v of 0.36 is, let us recall, the maximum threshold we have set for ourselves. By increasing this threshold further, we could increase the proportion of relevant articles (especially since for x=0.36, the recall has not yet started to decrease), but the number of articles selected would be too small to perform robust analyses (Table 2).

$$R = \frac{N_{FS,w2v}}{N_{FS}} \quad ; \quad P = \frac{N_{FS,w2v}}{N_{w2v}} \quad ; \quad F = 2 \times \frac{R \times P}{R + P} \tag{1}$$

where R, P, and F represent recall, precision, and F-measure, respectively. N_{FS} is the number of articles annotated as having the topic "food security", N_{w2v} is the number of articles classified by w2v as having the topic "food security", and $N_{FS,w2v}$ is the number of articles annotated and classified by w2v as having the topic "food security".

Table 3: Comparison of the distributions (in %) of the manually annotated classes ("0", "1" and "2") on 24 articles, as well as the error rates (in %), recall, precision and F-measure associated with the w2v thresholds compared to the manually annotated classes, for each threshold x = 0.28 ; 0.32 ; 0.36.

	x=0.28	**x=0.32**	**x=0.36**
Proportion of "0"	62.5	75	66.7
Proportion of "1"	37.5	20.8	20.8
Proportion of "2"	0	4.2	12.5
Error rate	62.5	50	41.7
Recall	0.33	0.5	0.63
Precision	0.25	0.25	0.42
F-measure	0.28	0.33	0.5

Manual annotation of texts is a time-consuming and concentration-intensive exercise, with guidelines that must be simple and unambiguous but often complex to apply. Given the time constraint imposed in this study, the small number of thresholds and numbers of articles tested here showed their limitations. Although the results obtained as well as the chosen threshold make sense, further work should be devoted to testing a larger number of thresholds with more associated articles, allowing, for example, setting the threshold more finely by using

proven criteria such as receiver operating characteristic (ROC) curves, which is a recognized tool to help in the choice of thresholds [2].

Our experiments allowed us to identify a threshold of 0.36 to determine texts related to the theme of food security conveyed in the articles. This threshold is applied to the phase **(1)** of the process. Then, we evaluate the negativity threshold to be applied for the **(2)** phase.

2. Validation of the threshold at which an article is considered negative

The relevance of the VADER model for measuring negativity in texts has been highlighted in the scientific literature [3]. We wish to evaluate how well specific thresholds associated with the VADER model predict negativity in the case of newspaper articles dealing with food security. These selected thresholds will then be applied in the process. We present here the methodology used to set an optimal negativity threshold in our context, from which an article is considered negative. The choice of this threshold is subject to the same dilemma between the relevance of the selected articles on the one hand and the quantity of articles available for the analysis on the other hand. The threshold was chosen with a methodology similar to the one used for the choice of the x threshold associated with w2v.

We first sort the articles by their negativity rate as calculated with the VADER model, restricting ourselves to articles with w2v scores higher than 0.36 (i.e., related to food security). Then, we construct an interval within which the final negativity threshold will be chosen. The lower bound is set empirically. By checking (reading) for several "threshold" values of approximately 30 articles whose negativity rate is just above the value, the value is chosen as the lower bound if at least a quarter of the checked articles are considered to be negative. We start with the value 0 and then advance in steps of 0.025. The first value tested for which we have at least a quarter of negative articles is 0.05, so we validate this value as the lower bound. For the selection of the upper bound, it is again the corresponding numbers of articles associated with each region studied that must hold our attention, for the same reasons as previously mentioned. Empirically, we consider an upper bound to be acceptable as long as at least 100 negative articles are associated with the corpus, and each region is associated with at least one negative article. For the choice of this threshold, the constraint on the number of articles retained is less strong than for the x threshold of w2v because the analysis allowed by this threshold is limited to calculations of average negativity rates (e.g., by regions or years). We check for several values for which these minimum constraints are verified on the corpus and for each region, starting with 0.05 (the lower bound) and advancing by steps of 0.025. Table 4 details the number of food security-related articles selected as negative

over the entire corpus and for each of the 3 regions studied as a function of the value chosen as the detection threshold (0.05, 0.075, 0.1 and 0.125). The last value tested for which we have a population size on the whole corpus of at least 100 as well as nonzero populations linked to each region is 0.1, so we choose this value as the upper bound.

Table 4: Illustration of the number of articles related to food security selected as negative on the whole corpus and for each of the 3 regions studied (Centre, Hauts-Bassins, Sahel) according to the chosen threshold value. The values in red correspond to the numbers that are too small for future analyses

Threshold	**Corpus**	**Centre**	**Hauts-Bassins**	**Sahel**
0.05	689	94	18	34
0.75	293	37	6	18
0.1	107	12	1	8
0.125	35	4	0	3

To validate the negativity threshold, we set 3 thresholds on the interval [0.05,0.1] that we constructed, 0.05, 0.075 and 0.1, which are evenly distributed between the two "limit" values of the interval. We then examine the extent to which each threshold discriminates relevant articles by checking whether articles with a negativity rate in the critical area of each threshold (i.e., just above) are negative in a sufficiently large proportion.

For each potential threshold, the 30 articles with the nearest and highest negativity rates above each of the thresholds were selected, and the negativity (i.e., the tendency of an article to discuss serious or worrisome topics using terms with negative connotations) of these articles was manually checked by an expert and annotated ("0": nonnegative article, "1": negative article). Our criterion for the choice of the threshold is that the threshold has in its critical zone a proportion of relevant articles (i.e., annotated as negative) significantly higher than for all the other thresholds and that the percentage of relevant articles in its critical zone is at least 50 %. We do not perform a recall/precision calculation here because we focus exclusively on the ability of the VADER model and the potential thresholds not to classify as negative too many articles that are not actually negative (false positives), which could bias the analyses performed on this group of selected articles. The number of 30 articles used by the threshold in this case is sufficient to obtain statistically significant results. We performed a χ^2 test of equality of the proportions of articles annotated as negative associated with the 3 thresholds, which is significant (p value < 0.05). The percentages of negative articles associated with each threshold illustrated in Table 5 are thus significantly different and consequently comparable. We can affirm that the threshold of negativity of 0.1 makes it possible to maximize the proportion of associated negative articles (67 %), which is higher than the "limit" proportion of 50 % fixed at the beginning. We therefore set the threshold at which an article is considered negative to 0.1.

Table 5: Percentage of articles manually labeled by an expert as negative for 3 groups consisting of the 30 articles with the closest negativity rates above the negativity thresholds 0.05, 0.075 and 0.1

Threshold	**Percentage of negative articles**
0.05	30%
0.075	47%
0.1	67%

For the w2v threshold that we set earlier, more in-depth studies that test more thresholds and articles would be desirable to clarify and reinforce our choices. Finally, it should be noted that the negativity threshold validated here is specific to our context, which is newspaper articles dealing with food security. This threshold is also particularly low; we believe that this is due to the type of textual domain, the newspaper article, which is bound to a duty of neutrality in the way it presents the news. For other textual media (e.g., scientific literature, NGO newsletters, and social network messages) and other themes whose thought patterns and writing styles may be very different, there is no guarantee that the optimal negativity thresholds are comparable to ours. This is the rationale for not using a large annotated corpus to set this threshold because in addition to there being very few in the French language, the few existing corpora (such as those proposed by DEFT[2]) are not adapted to our context.

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.

References

[1] J. L. Fleiss, J. Cohen, The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability, Educational and Psychological Measurement 33 (3) (1973) 613–619. doi:10.1177/001316447303300309.

[2] H. Delacour, A. Servonnet, La courbe roc (receiver operating characteristic): principes et principales applications en biologie clinique, in: Annales de biologie clinique, Vol. 63, 2005, pp. 145–154.

[3] C. H. E. Gilbert, 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.

[2] https://deft.limsi.fr/