Thematising online food risks: Comparison of a manual tagging procedure and topic modelling

Valentina Rizzoli[1*], Mirko Ruzza[1], Luca Lunardi[1], Barbara Tiozzo[1], Licia Ravarotto[1]

[1] Health Awareness and Communication Department – Istituto Zooprofilattico Sperimentale delle Venezie

[*] vrizzoli@izsvenezie.it

Abstract

One of the main needs to face in front of a huge amount of contents is to classify them in themes. The present study compares a manual tagging with an automatic procedure implemented in the context of Machine Learning applied to food risk issues. For a year, web sources have been monitored through the web monitoring application Web-Live®, developed by the company Extreme s.r.l. (http://www.web-live.it) and 12,163 contents were collected. Subsequently, the items were in parallel labelled according to two procedures: a manual (Elo & Kyngäs, 2008) and an automatic one (cf. Tuzzi, 2003), that is the Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003) implemented in the "topicmodels" package (Grün & Hornik, 2011) available in R. Discrepancies and overlapping of the labelling and the classification have been observed using the data visualisation software Qlik Sense®. Both procedures highlighted mostly the same contents as regards the labelling goal, and return a similar classification regarding the overlapping topics. The analysis of both outputs showed that the automatic procedure preferably returned precise and detailed topics, whereas the manual procedure enabled more levels of tagging. Results have been further discussed highlighting the criticality and potential of the approaches addressed, to inform any additional application.

Keywords: Content analyses, manual tagging, latent Dirichlet allocation, food risk communication

1. Introduction

Online resources (e.g., websites, social media) have become one of the main channels through which information about food risk is published and sought, thus contributing to building readers' perception and knowledge (Kuttschreuter et al., 2014). On the one hand, researchers are faced with a rich source of natural data that can provide information concerning the most debated issues and cases concerning food risks, on the other they face the challenge of managing and analysing these data, in order to successfully inform food risk communication by competent health authorities.

To answer the necessity to classify the contents it is possible to resort to the content analyses, that could be considered an umbrella term, and synonym of text mining or text analyses, i.e. the process of collecting, coding, analyse and interpret the information inherent in one or more texts, returning their content to a new form (Tuzzi, 2003; 2010; cf. Berelson, 1952). It could be classical/manual (as the thematic analysis; Flick 2009) or modern/automatic, i.e. based on a bag of words approach (Tuzzi, 2003). With the presence of an increasing amount of available content, but also of tools, various methods have been developed that allow implementing these analyses automatically. Among the methods that allow identifying thematic structures in collections of texts automatically, there is the latent Dirichlet allocation

(LDA), a probabilistic topic modelling algorithm developed in the context of machine learning (Blei, Ng, & Jordan, 2003; Blei, 2012). On the one hand, in front of a large amount of data, automated analyses are essential in terms of cost-efficiency. Moreover, they respond more easily to the reproducibility requirement. On the other hand, they run into the problem of the validity of the encoding, which is more easily overcome by a rigorous manual classification (cf. Scharkow, 2017).

The present study aims at comparing a manual tagging with the LDA, to reflect on the limits and potentialities of an automatic versus manual approach in the content analysis and to validate the goodness of the automatic tagging with respect to manual output. To this extent, manual analysis was preliminarily performed to understand the corpus; after that, the automatic procedure was run as well, and both outputs were compared and analysed for similarities and divergencies. Both procedures are applied to the specific context of food risk and safety issues.

2. Method

For a year, web sources (news media outlets, websites, blogs, forums, public social media accounts) have been monitored through the web monitoring application Web-Live®, developed by the company Extreme s.r.l. (http://www.web-live.it). It automatically retrieved relevant contents according to a monitoring profile related to food risk. A system of rules based on the combination of keywords and logical operators were used to query search engines (Google, Bing, Yahoo) and social network websites (Facebook, Twitter, Google+, YouTube, Instagram) to retrieve contents pertinent to food risk. Up to 50 contents per day, among those retrieved, were manually validated. At the end of the monitoring period, 12,163 contents were collected[1]. Subsequently, the items were in parallel labelled according to two procedures, a manual and an automatic one (cf. Tuzzi, 2003).

2.1. Manual tagging

As regards the manual labelling, according to an open coding process (Elo & Kyngäs, 2008), a label was assigned to each item of the corpus by two researchers separately using the spreadsheet Microsoft Excel. Each content has been checked and refined iteratively following a bottom-up process. If two or more items referred mainly to the same topic, they were assigned the same tag. Mutually exclusive labels were applied according to the prevalent theme treated. New tags were added to a list as they were created. These labels were thus grouped into broader ones. Researcher mostly agreed, even if with some divergences. The discrepancies were discussed and resolved until an agreement, and a third coder was involved with a supervisory role and guaranteed consistency in the tag assignment.

2.2. Latent Dirichlet allocation

As regards the automatic procedure, the corpus has been pre-processed with TaLTaC[2] (version, 2.10.2, Bolasco, Baiocchi, & Morrone, 2000; Bolasco, 2010) by reducing uppercase letters to lowercase. The lexicometric measures showed a good redundancy (Table 1).

[1] Only one item from Youtube was gathered. We decided to not consider it since it was the only one video content.

Table 1 – lexicometric measures of the corpus

N—Word-tokens	4,166,610
V—Word-types	82,872
(V/N)*100—Type/Token ratio	1.99
(VI/V)*100—Hapax percentage	39.43

Multiwords with frequency ≥ 80 have been individuated by means of an automatic information retrieval procedure that recognise repeated informative sequences of words (Pavone, 2018). The pre-processed corpus was thus exported. Topic detection procedure (Blei, Ng, & Jordan, 2003) latent Dirichlet allocation (LDA) implemented in the "topicmodels" package (Grün & Hornik, 2011) available in R was applied. According to the Griffiths e Steyvers (2004) model, that uses the log-likelihood variation, the suggested number of topics to individuate is approximately 56 (Figure 1). The LDA returns the most probable words for each topic and the association among topics and articles. As for the manual procedure, the authors assigned a label to the individuated topics, by observing the list of the most associated words to each topic.

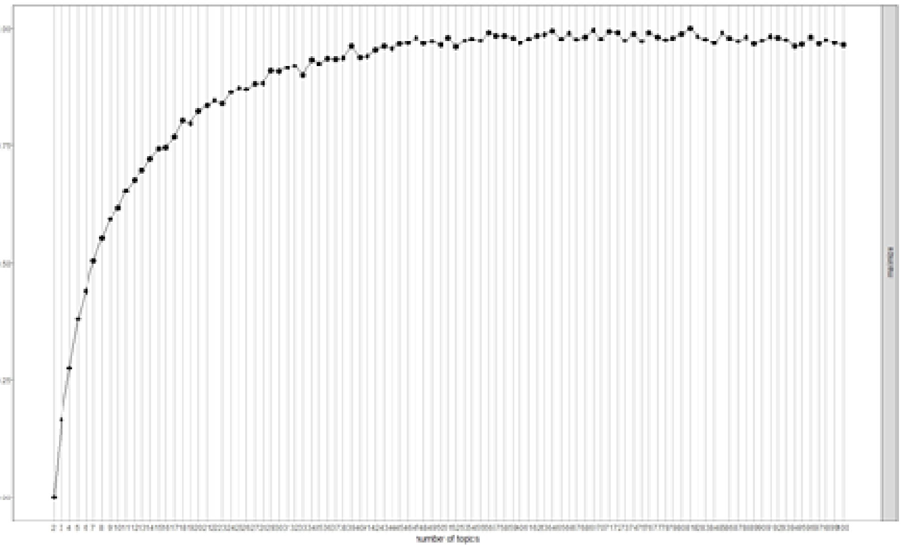


Figure 1- Log-likelihood for increasing numbers of topics (2-100)

To examine the effectiveness and validate the procedures, both the outputs were compared using the data visualisation software Qlik Sense®. Thanks to the ID associated with each item, the software allowed to observe which item was associated with which topic and to verify congruencies and discrepancies in the two classifications.

3. Results

3.1. Manual tagging

With the manual tagging, 45 categories and 10 macro-categories have been individuated (see Table 2 for the details). The macro-categories include contents mainly referring to one or more specific food risks and related control activities and alerts (e.g. *nutritional risks*, 21.6% of the total corpus; *outbreaks, controls and alerts*, 17.0%; *chemical risks*, 13.7%), and contents specifically mentioning food emergencies (*media cases*, 12.9%). Finally, other residual categories generally refer to food safety as a public health problem, without mentioning or focusing on specific risks (e.g., *production/economic aspects*, 12.6% of the total corpus).

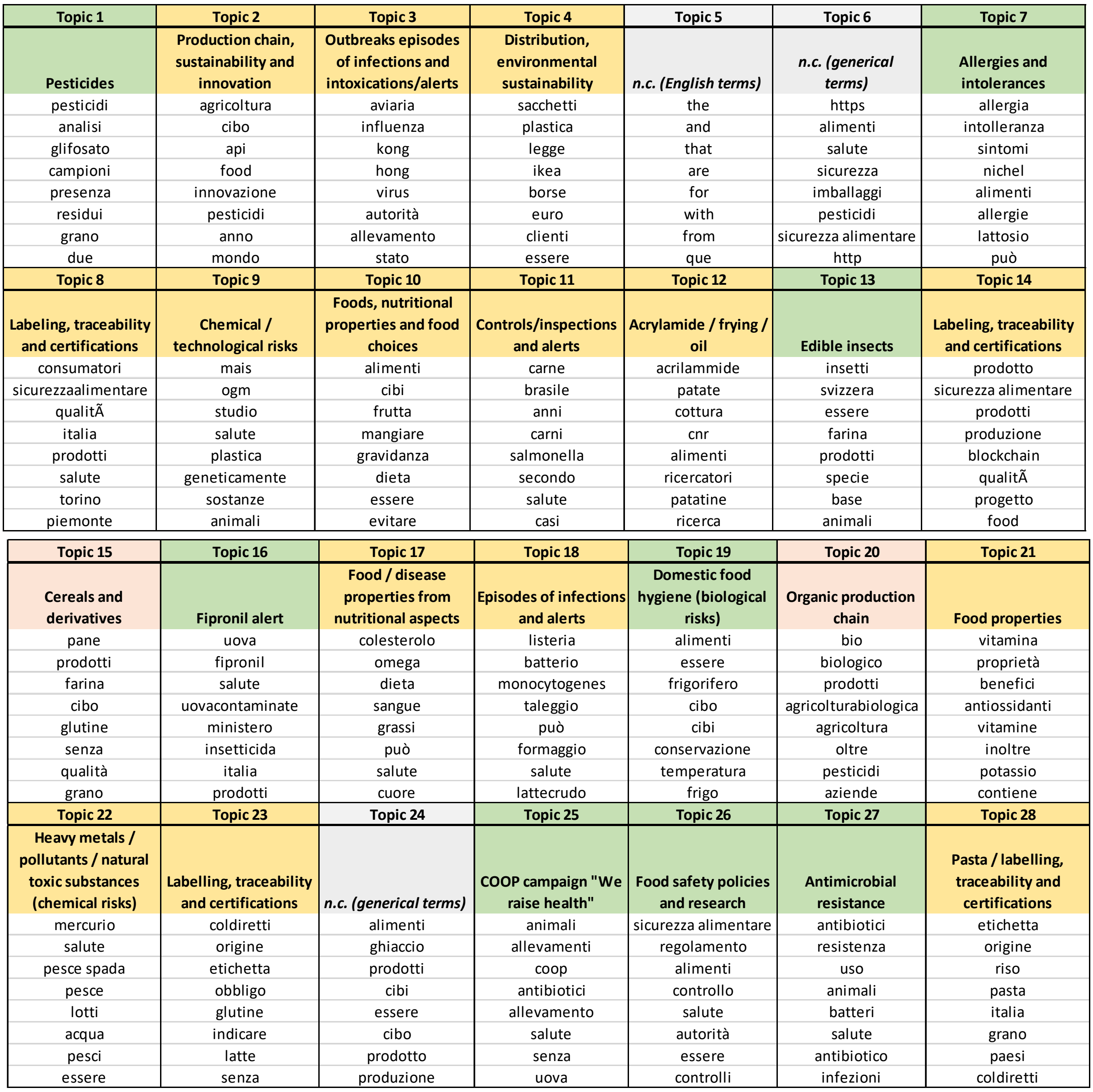
The macro-category *nutritional risks* includes practical advice for the consumer on the properties of foods, nutrients, diets or of specific eating habits. It includes categories as *beneficial/harmful properties of food and nutrients* (57.1% of the category), *allergies and intolerances* (16.7%), and *diseases related to nutritional risks* (15.6%), as shown in Table 2. The macro-category *chemical risks* (13.7% of the corpus) is divided into smaller categories mainly mentioning *pesticides and residue of phytosanitary treatments* (30.8% of the macro-category), *antibiotics and antimicrobial resistance* (19.1%), *additives* (15.9%), *food contact materials* (9.9%). The macro-category *media cases* (12.9% of the corpus) deals mainly with food alerts notifications as *fipronil alert* (31.5% of the contents of the category), *PFAS alert* (22.0%), and *glyphosate debate* (19.0%), and news on the measures adopted to cope with it. The 12.6% of the corpus has been classified in the macro-category *production/economic* aspects that contains articles dealing with origin and traceability of food products, the role of certifications, and issues related to labelling (e.g., *labelling, traceability and certification*s, 32.2%; *production chain and innovation*, 28.1%; *made in Italy/local products vs foreign products*, 19.3%). In the macro-category *biological risks* (6.1% of the corpus) the main arguments treated are *bacteria, viruses, and parasites* (48.5%) and *food hygiene at home* (37%). A smaller amount of items has been labelled as *Political and institutional aspects* (4.6% of the corpus) that concerns *food safety policies and research* (76.0%) and *official control of foodstuffs* (23.0%). The remaining articles concern *Risk of specific foods/situation,* as during pregnancy or child nutrition (28.9% of the macro-category) or on vacation (20.7%); *communication campaigns* 1.6% of the corpus (e.g., *"let's grow health"*, 60.6% of the macro-category); and other various aspects (6.2% of the total) as the *sustainability of the food production system* (36.6%) or *plant and animal diseases* (28.7%).

Table 2 – Macro-categories and categories identified with the manual content analysis. Number of manually classified contents: 12.163 (100%). Colours of lines respect the proportion of the categories inside the macro-categories.

SPECIFIC FOOD RISKS (76.6%)			
Nutritional risks (21.6% of the corpus)	Beneficial/harmful properties of food and nutrients (57.1%)	*Media cases (12.9% of the corpus)*	Fipronil alert (31.5%)
	Allergies and intolerances (16.7%)		PFAS alert (22%)
	Diseases related to nutritional risks (15.6%)		Glyphosate debate (19%)
	Habits, diets and food choices (10.6%)		Palm oil debate (7.4%)
Outbreaks, controls and alerts (17% of the corpus)	Withdrawals/recalls and alerts (52.3%)		CETA debate (7.3%)
	Inspections, seizures and penalty measures (41.8%)		Beef hormone dispute (5.9%)
	Episodes of infection or intoxication (6%)		Edible insects (4%)
Chemical risks (13.7% of the corpus)	Pesticides and residues of phytosanitary treatments (30.8%)		Salmonella in milk powder (3%)
	Antibiotics and antimicrobial resistance (19.1%)	*Biological risks (6.1% of the corpus)*	Bacteria, viruses and parasites (48.5%)
	Additives (15.9%)		Food hygiene at home (37%)
	Food contact materials (9.9%)		Food hygiene in the production chain (11%)
	Environmental pollutants (7.5%)		Water hygiene (3.5%)
	Natural toxic substances (6.1%)	*Risks of specific foods/situations (3.8% of the corpus)*	Specific risky foods (39.1%)
	Residues from the production process (5.9%)		Nutrition during pregnancy/feeding of children (28.9%)
	Substances produced by cooking (4.8%)		Eating during summer/on vacation (20.7%)
			Debunking of fake news (11.2%)
		Communication campings (1.6% of the corpus)	“Let’s grow health!” communication campaign (60.6%)
			“Dangerous foods blacklist” communication camp. (39.4%)
FOOD SAFETY IN GENERAL (23.4%)			
Production/economic aspects (12.6% of the corpus)	Labelling, traceability and certifications (32.2%)	*Political/institutional aspects (4.6% of the corpus)*	Official control of foodstuffs policies (77%)
	Production chain and innovation (28.1%)		Food safety policies and research (23%)
	Made in Italy/local products vs foreign products (19.3%)	*Other aspects (6.2% of the corpus)*	Sustainability of the food production system (36.6%)
	Distribution, trade and consumption (10.2%)		Events, anniversaries and dissemination (29.1%)
	Animal welfare (5.3%)		Plants and animal diseases (28.7%)
	Canteens and restaurants (4.9%)		Other (5.6%)

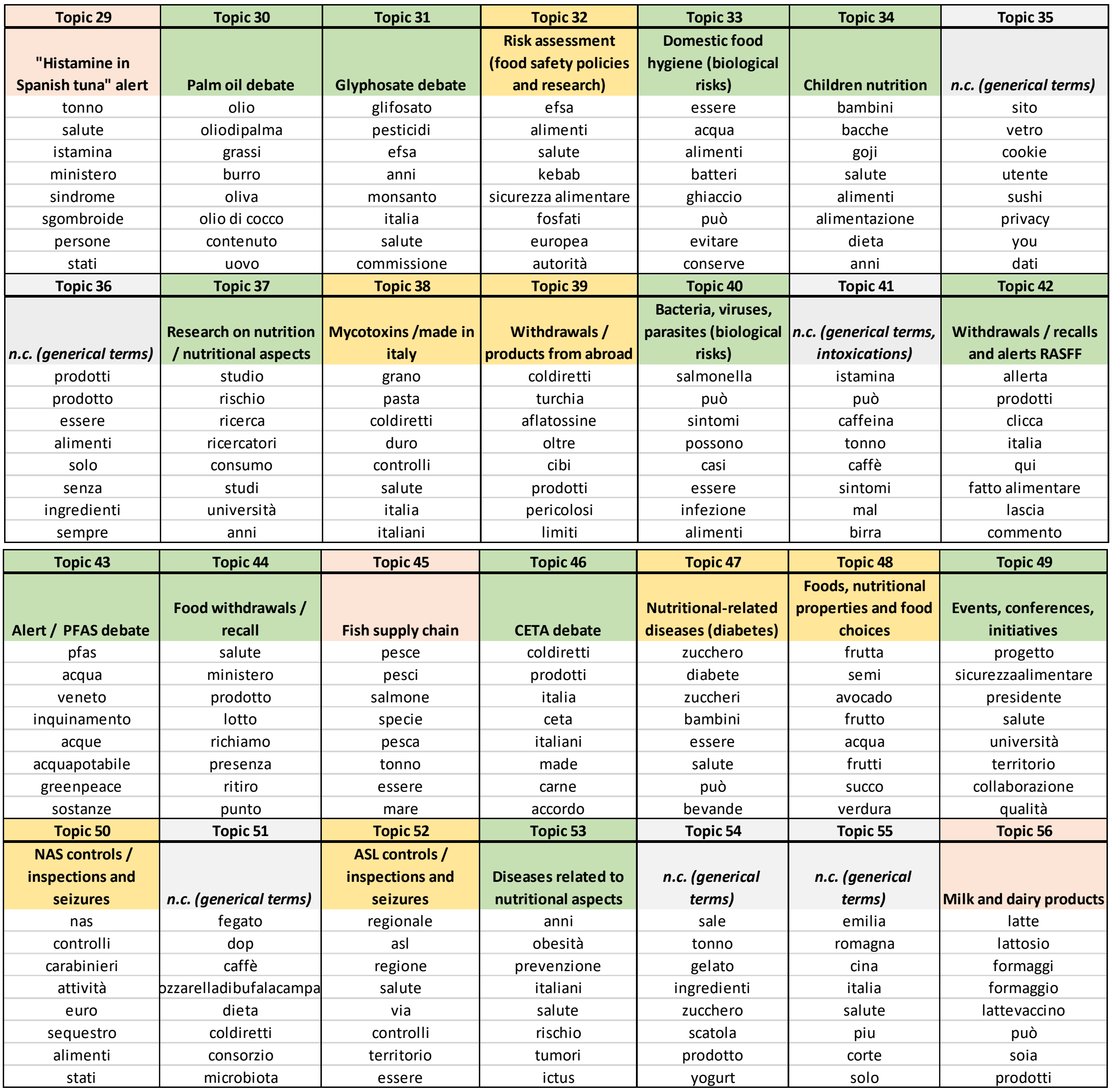
3.2. Latent Dirichlet allocation

The LDA returned 56 topics that have been labelled by the authors, as shown in Figure 2, according to the most probable words and articles associated with each one.



Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
Pesticides	**Production chain, sustainability and innovation**	**Outbreaks episodes of infections and intoxications/alerts**	**Distribution, environmental sustainability**	***n.c. (English terms)***	***n.c. (generical terms)***	**Allergies and intolerances**
pesticidi	agricoltura	aviaria	sacchetti	the	https	allergia
analisi	cibo	influenza	plastica	and	alimenti	intolleranza
glifosato	api	kong	legge	that	salute	sintomi
campioni	food	hong	ikea	are	sicurezza	nichel
presenza	innovazione	virus	borse	for	imballaggi	alimenti
residui	pesticidi	autorità	euro	with	pesticidi	allergie
grano	anno	allevamento	clienti	from	sicurezza alimentare	lattosio
due	mondo	stato	essere	que	http	può
Topic 8	**Topic 9**	**Topic 10**	**Topic 11**	**Topic 12**	**Topic 13**	**Topic 14**
Labeling, traceability and certifications	**Chemical / technological risks**	**Foods, nutritional properties and food choices**	**Controls/inspections and alerts**	**Acrylamide / frying / oil**	**Edible insects**	**Labeling, traceability and certifications**
consumatori	mais	alimenti	carne	acrilammide	insetti	prodotto
sicurezzaalimentare	ogm	cibi	brasile	patate	svizzera	sicurezza alimentare
qualitÃ	studio	frutta	anni	cottura	essere	prodotti
italia	salute	mangiare	carni	cnr	farina	produzione
prodotti	plastica	gravidanza	salmonella	alimenti	prodotti	blockchain
salute	geneticamente	dieta	secondo	ricercatori	specie	qualitÃ
torino	sostanze	essere	salute	patatine	base	progetto
piemonte	animali	evitare	casi	ricerca	animali	food

Topic 15	Topic 16	Topic 17	Topic 18	Topic 19	Topic 20	Topic 21
Cereals and derivatives	**Fipronil alert**	**Food / disease properties from nutritional aspects**	**Episodes of infections and alerts**	**Domestic food hygiene (biological risks)**	**Organic production chain**	**Food properties**
pane	uova	colesterolo	listeria	alimenti	bio	vitamina
prodotti	fipronil	omega	batterio	essere	biologico	proprietà
farina	salute	dieta	monocytogenes	frigorifero	prodotti	benefici
cibo	uovacontaminate	sangue	taleggio	cibo	agricolturabiologica	antiossidanti
glutine	ministero	grassi	può	cibi	agricoltura	vitamine
senza	insetticida	può	formaggio	conservazione	oltre	inoltre
qualità	italia	salute	salute	temperatura	pesticidi	potassio
grano	prodotti	cuore	lattecrudo	frigo	aziende	contiene
Topic 22	**Topic 23**	**Topic 24**	**Topic 25**	**Topic 26**	**Topic 27**	**Topic 28**
Heavy metals / pollutants / natural toxic substances (chemical risks)	**Labelling, traceability and certifications**	***n.c. (generical terms)***	**COOP campaign "We raise health"**	**Food safety policies and research**	**Antimicrobial resistance**	**Pasta / labelling, traceability and certifications**
mercurio	coldiretti	alimenti	animali	sicurezza alimentare	antibiotici	etichetta
salute	origine	ghiaccio	allevamenti	regolamento	resistenza	origine
pesce spada	etichetta	prodotti	coop	alimenti	uso	riso
pesce	obbligo	cibi	antibiotici	controllo	animali	pasta
lotti	glutine	essere	allevamento	salute	batteri	italia
acqua	indicare	cibo	salute	autorità	salute	grano
pesci	latte	prodotto	senza	essere	antibiotico	paesi
essere	senza	produzione	uova	controlli	infezioni	coldiretti



Topic 29	Topic 30	Topic 31	Topic 32	Topic 33	Topic 34	Topic 35
"Histamine in Spanish tuna" alert	Palm oil debate	Glyphosate debate	Risk assessment (food safety policies and research)	Domestic food hygiene (biological risks)	Children nutrition	*n.c. (generical terms)*
tonno	olio	glifosato	efsa	essere	bambini	sito
salute	oliodipalma	pesticidi	alimenti	acqua	bacche	vetro
istamina	grassi	efsa	salute	alimenti	goji	cookie
ministero	burro	anni	kebab	batteri	salute	utente
sindrome	oliva	monsanto	sicurezza alimentare	ghiaccio	alimenti	sushi
sgombroide	olio di cocco	italia	fosfati	può	alimentazione	privacy
persone	contenuto	salute	europea	evitare	dieta	you
stati	uovo	commissione	autorità	conserve	anni	dati
Topic 36	**Topic 37**	**Topic 38**	**Topic 39**	**Topic 40**	**Topic 41**	**Topic 42**
n.c. (generical terms)	Research on nutrition / nutritional aspects	Mycotoxins /made in italy	Withdrawals / products from abroad	Bacteria, viruses, parasites (biological risks)	*n.c. (generical terms, intoxications)*	Withdrawals / recalls and alerts RASFF
prodotti	studio	grano	coldiretti	salmonella	istamina	allerta
prodotto	rischio	pasta	turchia	può	può	prodotti
essere	ricerca	coldiretti	aflatossine	sintomi	caffeina	clicca
alimenti	ricercatori	duro	oltre	possono	tonno	italia
solo	consumo	controlli	cibi	casi	caffè	qui
senza	studi	salute	prodotti	essere	sintomi	fatto alimentare
ingredienti	università	italia	pericolosi	infezione	mal	lascia
sempre	anni	italiani	limiti	alimenti	birra	commento

Topic 43	Topic 44	Topic 45	Topic 46	Topic 47	Topic 48	Topic 49
Alert / PFAS debate	Food withdrawals / recall	Fish supply chain	CETA debate	Nutritional-related diseases (diabetes)	Foods, nutritional properties and food choices	Events, conferences, initiatives
pfas	salute	pesce	coldiretti	zucchero	frutta	progetto
acqua	ministero	pesci	prodotti	diabete	semi	sicurezzaalimentare
veneto	prodotto	salmone	italia	zuccheri	avocado	presidente
inquinamento	lotto	specie	ceta	bambini	frutto	salute
acque	richiamo	pesca	italiani	essere	acqua	università
acquapotabile	presenza	tonno	made	salute	frutti	territorio
greenpeace	ritiro	essere	carne	può	succo	collaborazione
sostanze	punto	mare	accordo	bevande	verdura	qualità
Topic 50	**Topic 51**	**Topic 52**	**Topic 53**	**Topic 54**	**Topic 55**	**Topic 56**
NAS controls / inspections and seizures	*n.c. (generical terms)*	ASL controls / inspections and seizures	Diseases related to nutritional aspects	*n.c. (generical terms)*	*n.c. (generical terms)*	Milk and dairy products
nas	fegato	regionale	anni	sale	emilia	latte
controlli	dop	asl	obesità	tonno	romagna	lattosio
carabinieri	caffè	regione	prevenzione	gelato	cina	formaggi
attività	ozzarelladibufalacampa	salute	italiani	ingredienti	italia	formaggio
euro	dieta	via	salute	zucchero	salute	lattevaccino
sequestro	coldiretti	controlli	rischio	scatola	piu	può
alimenti	consorzio	territorio	tumori	prodotto	corte	soia
stati	microbiota	essere	ictus	yogurt	solo	prodotti

Figure 2 – Labelled topics with the most probable words associated to them. In green those that are entirely overlapping, in yellow partially overlapping, in red not overlapping. Grey ones are not classifiable.

Nine topics have not been classified (grey in Figure 2), since they are given by too generic terms, or they identified and gathered specific aspects related to the language (e.g., topic 5 includes only English terms). Twenty topics out of 56 (topics 1, 7, 13, 16, 19, 25, 26, 27, 30, 31, 33, 34, 37, 40, 42, 43, 44, 46, 49, and 53; green in Figure 2) are completely overlapping with the categories individuated in the manual tagging. Twenty-two topics (2, 3, 4, 8, 9, 10, 11, 12, 14, 17, 18, 21, 22, 23, 28, 32, 38, 39, 47, 48, 50, and 52; yellow in Figure 2) have been individuated in the manual tagging, but they are part of broader topics, or they contain more than one topic among those individuated in the manual tagging. For example, as regards the former case, topics 50 and 52, concerning controls respectively from NAS (acronym for Nuclei Antisofisticazioni e Sanità, one of the Italian health authorities in charge of inspections and controls of the food chain) or ASL (acronym for Aziende Sanitarie Locali, Italian local health authorities), are both incorporated in the *inspections, seizures and penalty measures* category of the manual labelling. As regards the latter case, topic 12, concerning Acrylamide/frying/oil, refers to both the manual individuated categories *substance produced by cooking* and *plant and animal diseases*. Five topics (15, 20, 29, 45, and 56; red in Figure 2) are not overlapping with the manually identified categories.

Then we moved to compare the classification of the articles. First, we compare the number of items belonging to the completely overlapping topics classified in both the procedures (Figure 3)[2]. Twelve labelled topics out of 17 (topics 34, 13, 19, 27, 43, 16, 31, 1, 7, 49, 26, 53) contain a difference of items that corresponds to less than half of those classified[3]. Then, we observed the proportion of the items classified with the same label in the manual procedure respect to the automatic ones (Figure 4). Topic 1, *pesticides*, includes 448 items, but most of them (119) have been manually classified into the manually individuated category *beneficial/harmful properties of food and nutrients*, only 98 (21.9%) in the *pesticides and residue of phytosanitary treatments,* that would be the corresponding one. Similarly, topic 34, *children nutrition*, and 49, *events, conferences, initiatives*, contain the same items classified manually with the corresponding label but they are not the most representative ones (respectively 3.8% and 19.4%). In topic 13, labelled as *edible insects*, there are only 63 items classified in the manual procedure as *edible insects* out of 150 items classified within the automatic one (42%), but it is still the most representative label. The same goes for topic 19, *domestic food hygiene (biological risks)* (48.7%), topic 25, *the "Let's grow health" communication campaign* (44.8%), topic 26, *food safety policies and research* (18.7%), topic 31, *glyphosate* (65.5%), topic 37, *research on nutrition / nutritional aspects* (61.2%), topic 46, *CETA debate* (31.4%), and topic 53, *diseases related to nutritional aspects* (54%). Topic 7, *allergies and intolerances* (86.8%), 16, *fipronil alert* (91.7%), 27, *antibiotics* (79.3%), 42, *withdrawals / recalls and alerts RASFF* (90.3%), 43, *PFAS alert* (80.6%), 44, *food withdrawals/recall* (96.7%), largely coincide (respectively 317 classified in the manual procedure compared to 275 in the automatic one, 399 compared to 435, 218 compared to 275, 167 compared to 185, 320 compared to 397, 351 compared to 363).

[2] Topics 30, 33, and 40 have not been compared.

[3] It is important to remember that, unlike manual classification, automatic ones can attribute multiple labels to the same item.

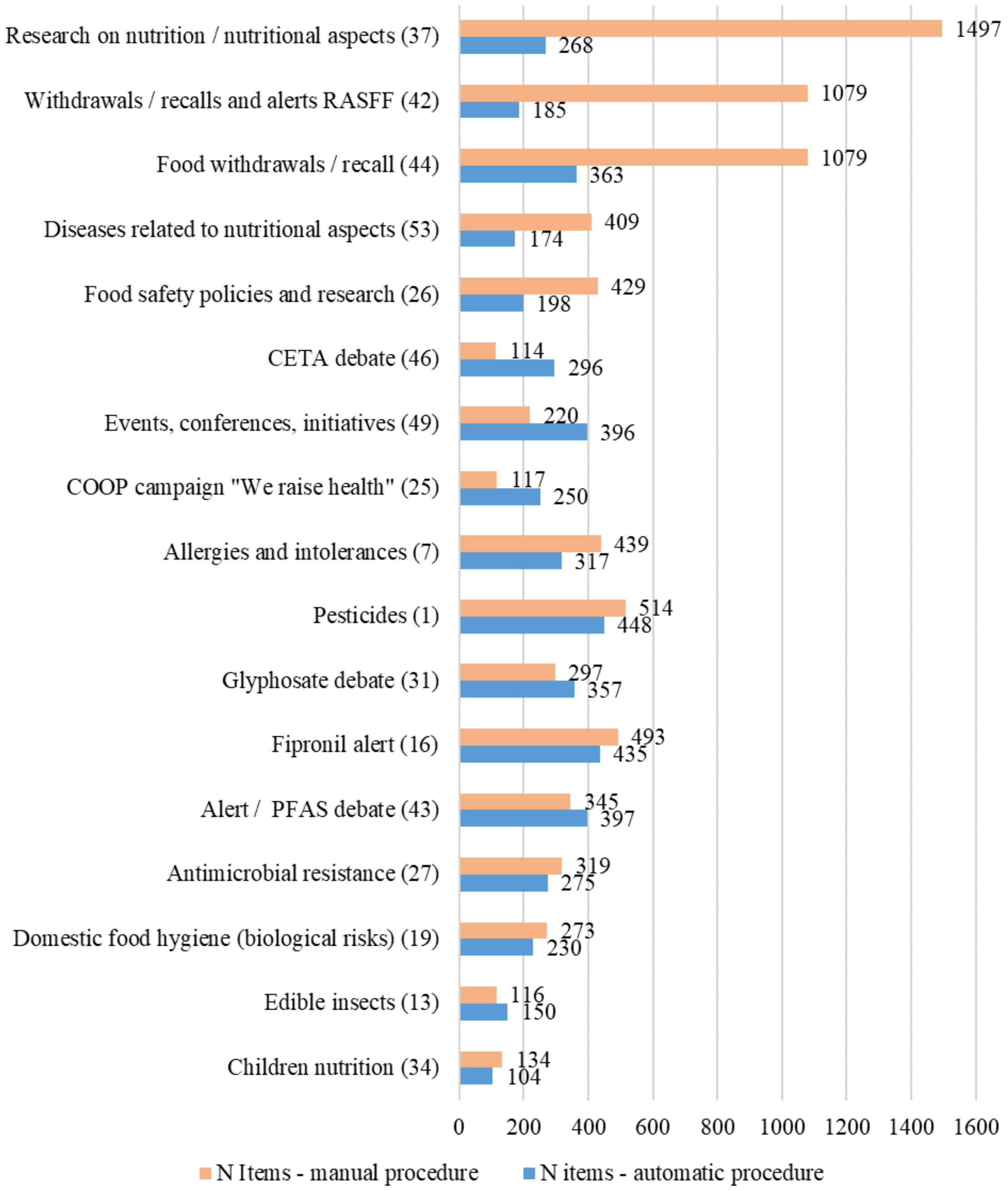


Figure 3 – Number of items contained in the topics with the completely overlapping labels with the manual procedure and the automatic one.

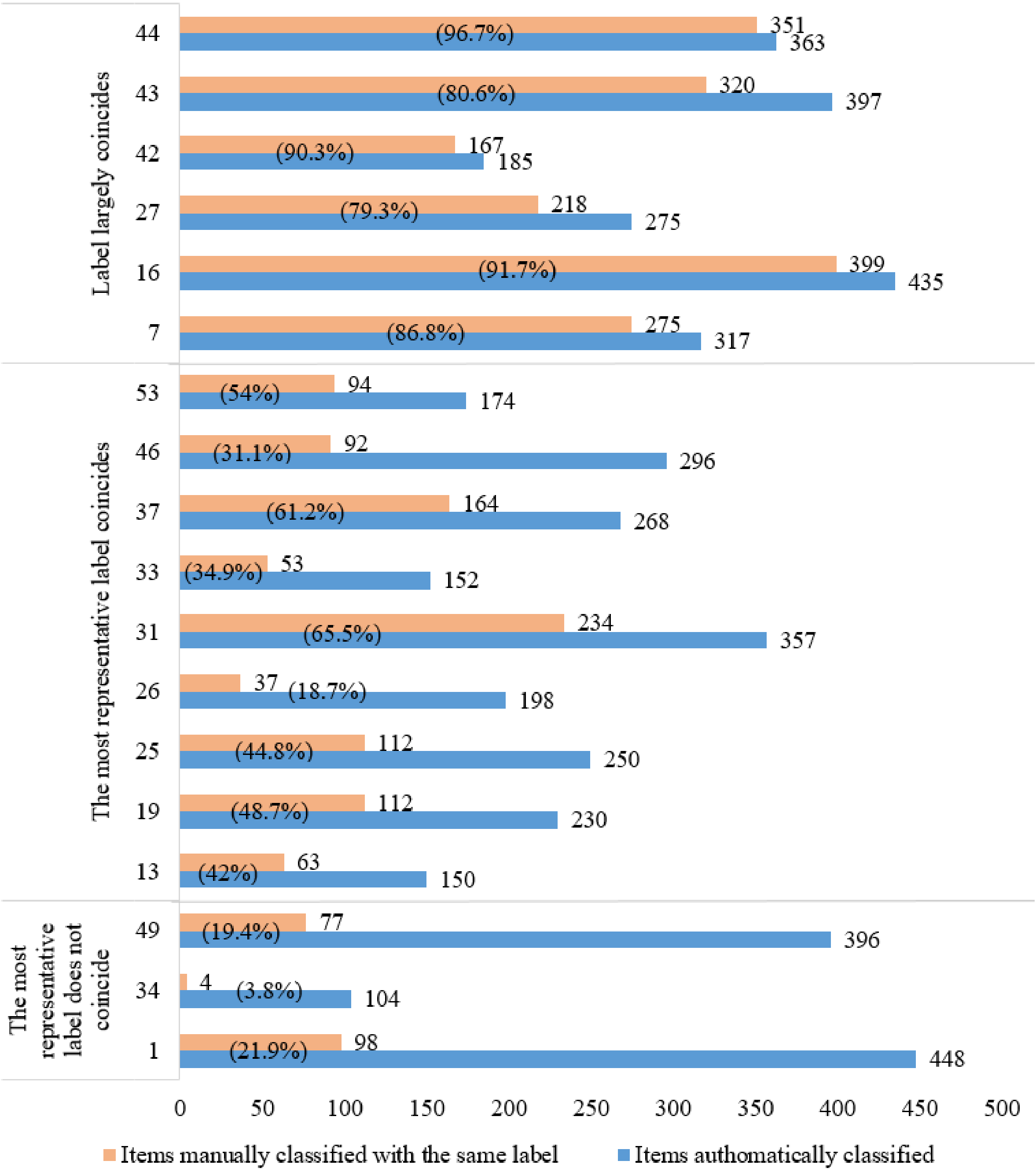


Figure 4 – Proportion of the items classified with the same label in the manual procedure respect to the automatic ones.

4. Discussion and Conclusions

In this paper two procedures of content analysis, one manual and one automatic (based on machine learning), were applied to understand and describe online contents related to food safety and related issues to evaluate to what extent the automatic procedure could mine and reveal meaning from the selected texts, possibly confirming the manual output.

To this extent, after performing manual and then automatic analysis, we first compared the topics individuated by both procedures. Most of them overlapped entirely or partially (42 out of 56) respect to the labels. Both procedures highlighted the presence of "media events" that occurred during the reference period, such as the "fipronil alert", the "PFAS alert" and the debate on the use of glyphosate; the development of communication campaigns by private companies and institutions to promote a shared "food safety education"; the presence of recurrent or ongoing topics that mainly refer to risk/benefit of foodstuffs and the sustainability of food safety policies. In general, the automatic procedure preferably returned precise and detailed topics, whereas the manual one enabled more levels of tagging, ranging from a general overview to an in-depth characterisation of online representation of food risks. Main differences are attributable to classification criteria not considered useful in manual tagging (e.g., specific foods related topics as cereals, topic 5, or milk and dairy products, topic 56). In this case, in fact, the manual groupings of categories took place based on the type of risk rather than the type of food, due to the choice of creating mutually exclusive categories. Other differences are ascribable to the algorithm functioning. For example, topic 12 put together two topics *substance produced by cooking* and *plant and animal disease* because they have in common the recurrence words as *oil.* In fact, acrylamide as a product of frying and xylella (disease of olive trees) have the oil in common. What just explained underlines, on the one hand, the importance of in-depth knowledge of the topic by the researchers; otherwise, it would have been difficult to grasp the interpretation mentioned above. On the other hand, the importance of fully knowing the criterion with which the method works was highlighted, to understand if the best approach has been used with respect to the purposes.

Then, we preliminary tried to compare the achievement of the classification aim. Most of the items classified within the topic automatically identified correspond to the ones in the manual classification (13 out of 20; three topics have not been compared, and the remaining four contain some of the same items categorised with the corresponding label, but they are not the most present). The expected result would be that the automatically identified topics contained more items classified within them than the corresponding manual categories, since, differently from the LDA, the manual categorisation produced mutually exclusive labels. However, we noticed that it is not always accurate, and this could be due to the higher level of specificity of the automatic categorisation respect to the manual one. On the other hand, by observing the proportion of items classified with the same label in the manual procedure, respect to the automatic one, the result is satisfying. With the automatic procedure, a good understanding of the text contents is provided. It does not replace a manual reading and remains a difference in the label assignment criterion, but it can be considered an efficient method for content analysis. This part of the work should be deepened; moreover, it would also be necessary to compare the classifications within the topics that are not entirely overlapping and to observe in detail the items that do not match in the classification.

To conclude with respect to the applied theme, that is food risk, both procedures provided an in-depth characterisation of online representation. This allows considering valid both the procedures, without forgetting the adequacy criterion with respect to the instruments (e.g., the amount of material) and the objectives (e.g., knowing how the method work to understand if it is the optimal one).

References

Berelson, B. (1952). *Content analysis in communication research.* Glencoe: The Free Press.

Blei, D. M., Ng, A. Y., & Jordan, M. (2003). Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, 3, 993–1022.

Blei, D. M. (2012). Topic modeling and digital humanities. *Journal of Digital Humanities*, 2(1).

http://journalofdigitalhumanities.org/2-1/topic-modeling-and-digital-humanities-by-david-m-blei/

Bolasco, S., Baiocchi, F., & Morrone, A. (2000). *TaLTaC^2: Trattamento automatico Lessicale e Testuale per l'analisi del Contenuto di un Corpus* [Computer software]. http://www.taltac.it/it/index.shtml

Bolasco, S. (2010). *Taltac 2.10. Sviluppi, esperienze ed elementi essenziali di analisi automatica dei testi.* Milano: LED.

Elo, S., & Kyngas, H. (2008). The qualitative content analysis process. *Journal of advanced nursing*, 62, 107-115.

Griffiths, T., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, 101(Supplement 1), 5228–5235

Grün, B., & Hornik, K. (2011). Topicmodels: An R package for fitting topic model. *Journal of Statistical Software*, 40(13), 1–30.

Kuttschreuter, M., Rutsaert, P., Hilverda, F., Regan, Á., Barnett, J., & Verbeke, W. (2014). Seeking information about food-related risks: The contribution of social media. *Food quality and preference*, 37, 10-18.

Pavone, P. (2018). Automatic Multiword Identification in a Specialist Corpus. In A. Tuzzi (Ed.), *Tracing the Life-Cycle of Ideas in the Humanities and Social Sciences* (pp. 151-166). New York: Springer.

Scharkow, M. (2017). Content analysis, automatic. *The International Encyclopedia of Communication Research Methods*, 1-14.

Tuzzi, A. (2003). *L'analisi del contenuto: introduzione ai metodi e alle tecniche di ricerca.* Roma: Carrocci.