Analysis of sustainability disclosure of food retailers in Europe

Johannes W.H. van der Waal

14-3-2022

Objective

- objectively chart the sustainability themes of food retailers
- identify relevant topics that we must not miss
- match our sustainability focus areas with those of retailers

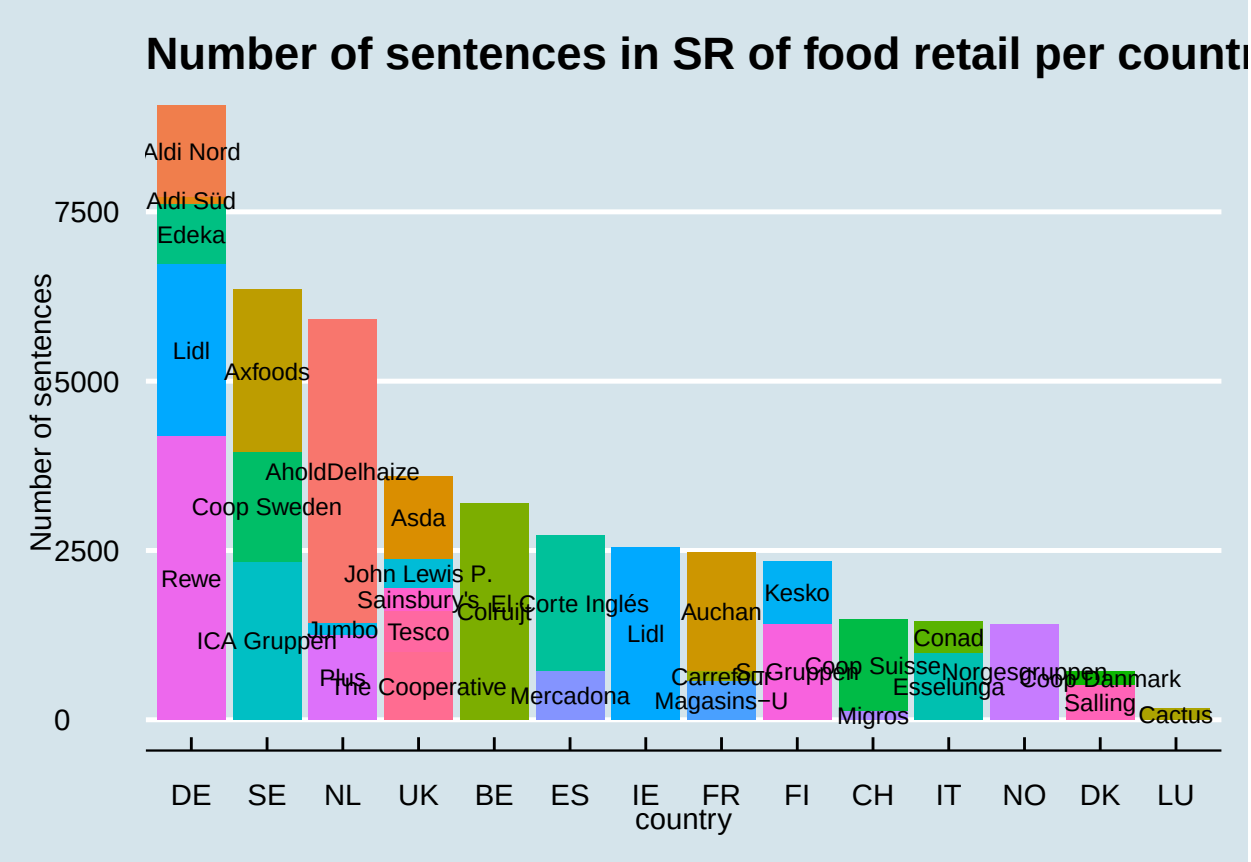
Method

Using *quanteda*, we perform text analysis of sustainability reports of European food retailers over 2020 (published 2021) or the closest year we can get. Some of the reports are in other languages than English. These were machine-translated into English using Deepl.com. Then we perform some supervised and non-supervised text analyses. Some of the sustainability reports were annual reports with a section on sustainability, so called "integrated" reports. OF all reports the document variables of the retailer publishing, the year of publication and the type (SR = sustainability report, AR = annual report, GC = Global Compact communication) were recorded.

The documents were converted from pdf to text and stored in a corpus (library or collection of documents). The corpus was further preprocessed, by removing numbers, spaces, various frequently appearing words, such as the names of the retailers, lower cased, stopwords were removed, and finally stemmed. The tokenized corpus was converted to a document-feature matrix (dfm). The dfm is the basis for further analysis.

Description of the reports

The following graph shows the number of sentences in the reports per retailer.



Topic Modeling We can perform an unsupervised topic modeling procedure on the corpus using Latent Dirichlet Allocation. This allows to automatically extract a number of topics accross documents.

Latent Dirichlet Allocation (LDA)

LDA assumes that every document in a corpus is a random mixture of latent topics. Every topic is considered as a mixture or distribution of words. LDA is an algorithm that tries to find the mixture of words that best defines a set number of topics, while at the same time estimating the mixture of topics that describes a document. LDA computes for every word the probability that it belongs to a certain topic (beta) and the estimated probability that a word in a document belongs to a certain topic (theta). Retrieving the words with the highest thetas allows to characterize the topic.

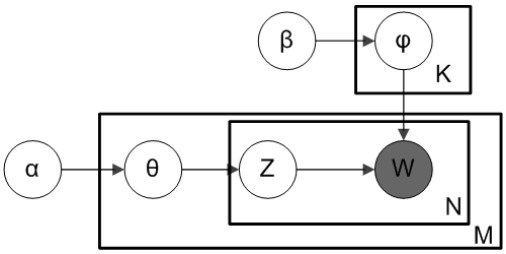


Figure 1: LDA model

In the above figure of the LDA algorithm ("plate model") denote:

- alpha - the per-document topic distributions,
- beta - the per-topic word distribution,
- theta - the topic distribution for document m,
- psi - the word distribution for topic k,
- z - the topic for the n-th word in document m, and
- w - the specific word

Applying LDA to the food retailer sustainability report corpus

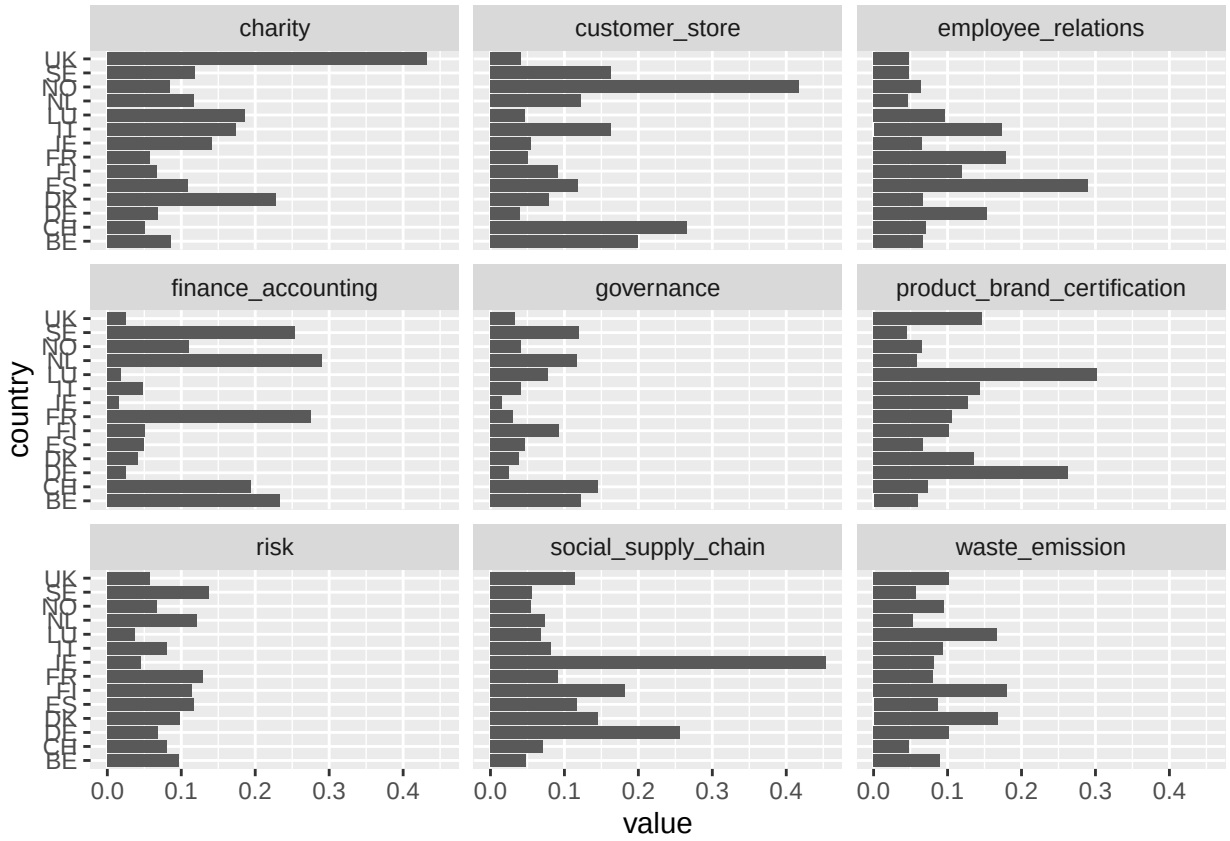
Using the package *seededlda* an LDA model was estimated, using 9 topics. This number was chosen somewhat arbitrarily but appeared to give a reasonable resolution and discrimination between topics.

The following table shows the ten words with the highest theta's per topic. These characterize the topic.

From the topic words we can derive a more descriptive name for the topic. Then we can plot the topics against the retailers. It is to note that the theta's of all the topics together sum to 1. So whether the retailer in question publishes a large report or a small one, the theta's give a relative probability of the presence of the topics in the document, irrespective of the document size. So there is no weighting for document size or the importance that subjects take relative to other documents.

country	retailer	no_of_tokens
NL	AholdDelhaize	78595
BE	Colruijt	48891
SE	ICA Gruppen	42485
SE	Axfoods	39928
ES	El Corte Inglés	28278
FI	S-Gruppen	23788
UK	The Cooperative	23721
NO	Norgesgruppen	22572
IT	Esselunga	21781
FR	Auchan	21102
UK	Asda	20336
CH	Coop Suisse	19686
FI	Kesko	16251
FR	Magasins-U	15804
ES	Mercadona	14074
NL	Plus	12419
DK	Salling	11059
IT	Conad	9991
IE	Lidl	9195
UK	John Lewis P.	8215
DE	Aldi Nord	4424
UK	Tesco	3340
UK	Sainsbury's	3230
DK	Coop Danmark	2766
LU	Cactus	2510
NL	Jumbo	1855
DE	Aldi Süd	1215

topic1	topic2	topic3	topic4	topic5	topic6	topic7	topic8	topic9
wast	employe	product	financi	report	support	sustain	store	board
emiss	work	packag	asset	risk	food	supplier	custom	member
energi	train	sustain	tax	audit	communiti	chain	sale	share
store	manag	plastic	incom	inform	peopl	suppli	product	committe
reduc	compani	label	valu	manag	help	respons	market	manag
consumpt	health	food	liabil	financi	work	product	retail	compani
climat	develop	recycl	cash	statement	custom	gri	food	director
food	safeti	organ	net	data	make	right	increas	execut
scope	employ	anim	amount	control	donat	social	onlin	meet
target	total	certifi	statement	intern	local	human	shop	remuner



Likewise, we can check if there are marked differences in topics between types of reports, notably integrated or annual reports and sustainability reports.

If we had a time series of sustainability reports, it would be possible to visualize temporal changes in topics.