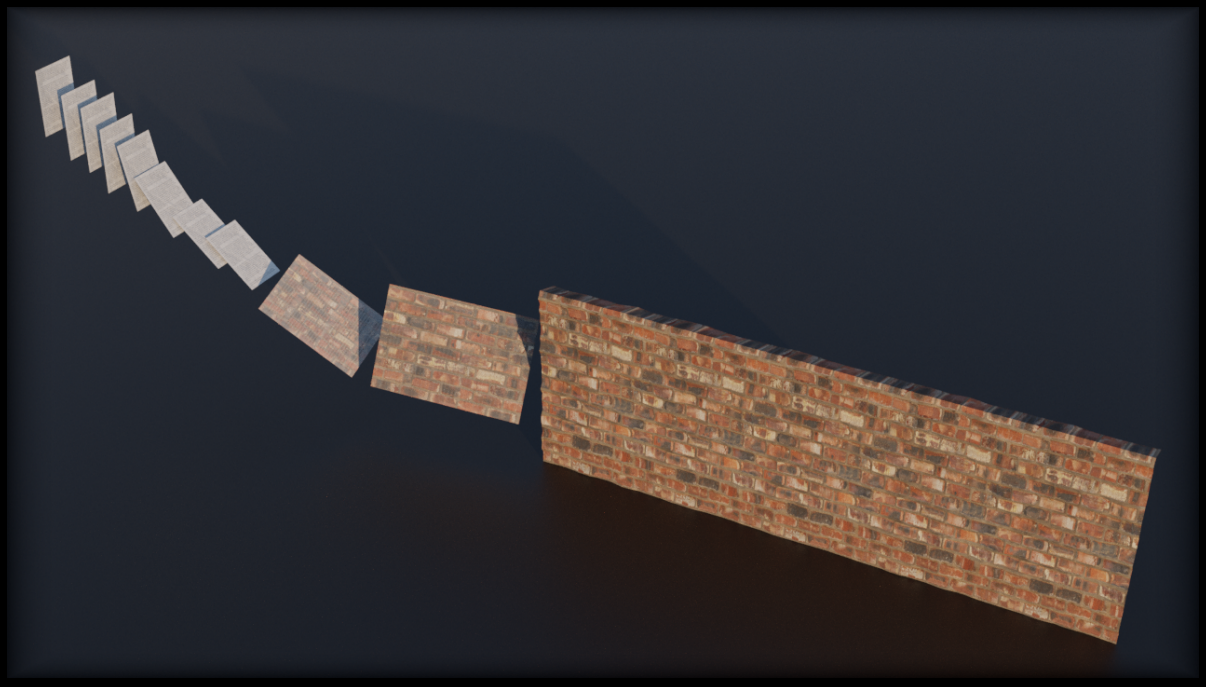
Modelling datafication of borders using public procurement documents



Mathias Kirkeng

Layout: typeset by the author using LaTeX.
Cover illustration: Mathias Kirkeng

Modelling datafication of borders using public procurement documents

Mathias Kirkeng
11904836

Bachelor thesis
Credits: 18 EC

Bachelor *Kunstmatige Intelligentie*

University of Amsterdam
Faculty of Science
Science Park 904
1098 XH Amsterdam

Supervisor
Prof. dr. T. Blanke

Institute for Logic, Language and Computation
Faculty of Humanities
University of Amsterdam
Spui 21
1012 WX Amsterdam

June, 2021

Abstract

Borders in the European Union (EU) are becoming increasingly datafied: that is, more data is collected and processed for migration management. This process is criticized on multiple issues in the social science literature. However, the agency tasked with managing borders in the EU, Frontex, has been criticized for its lack of transparency. Nevertheless, there is some information Frontex is compelled to publish, that being public procurement documents. This paper aims to analyse these documents using Latent Dirichlet Allocation (LDA), a topic modelling method. In this research, LDA is used to analyse datasets made up of contract award notices (CAN) and call for tender (CFT) documents. The resulting topics and timelines of topics give an empirical account of the datafication of borders and support the social science literature on this issue.

Keywords: borders, migration, datafication, latent dirichlet allocation, machine learning, topic modelling

Contents

Chapter 1

Introduction

Borders in the European Union (EU) have become increasingly datafied. Datafication is characterized by the increased collection and processing of large amounts of data related to migration for purposes of migration management. This is aimed at "visualising, registering, mapping and profiling mobile (sub)populations" (Broeders and Dijstelbloem, 2015). The increase in datafication is exemplified by the creation of systems such as EUROSUR, the European Border Surveillance System used to track illegal immigration into the EU (Broeders and Dijstelbloem, 2015) as well as migration-related databases such as EURODAC, SIS II, and VIS (Stenum, 2017). Another example is the increased usage of dual-use drones for border surveillance (Csernatoni, 2018).

1.1 Objections to datafication

This increased use of data for border control has been criticized for multiple reasons in the social science literature. Taylor and Meissner (2020), for example, describe the marketization of big data collection and analysis for border control and its implications. They note how this big data approach requires the formalisation of what a migrant is, what risk is and what forms of mobility are relevant. This is done with proxies in the available data. Due to the nature of the competitive market, the proxies chosen will be the ones most profitable which are the proxies that reinforce the crisis framing of migration that was dominant around 2015. Taylor and Meissner also argue that this crisis framing tends toward solutionism, meaning starting with a solution and defining the problem in a way that fits this solution. Furthermore, the authors highlight that this approach is based on the idea that the complex phenomenon of migration can be simplified and that, through the use of big data, human mobility can be predicted. The authors dispute this by pointing to empirical evidence to the contrary.

Similarly, Metcalfe and Dencik (2019) describe migrants as subjects of data experimentation performed by governmental, humanitarian and private actors. The authors also highlight that even when data is used for humanitarian reasons, those uses can be co-opted for migration management. An example of this is the use of cash cards issued by non-governmental organisations in Greece. These cards provide monetary aid to refugees but can be rendered unusable when a migrant leaves a certain geographical area. Metcalfe and Dencik argue that in this example cash cards are used to enforce a policy of migrant hotspots which goes against the 1951 refugee convention. In addition, the authors note the externalisation of borders through the collection of data outside of the border territory. The authors argue that this moves the border outward from their physical boundaries. This externalisation is exemplified by the use of drones above the mediterranean sea to preemptively deny access to migrants even if they might have a legal cause for refuge in the EU. Echoing Talyor and Meissner's argument of the complexity of migration, Metcalfe and Dencik also point out that datafication is based on the assumption that big data can be used to predict individual behaviour. The idea that one can act on those predictions to avoid threats then justifies the collection of an infinite amount of data. Finally, Metcalfe and Dencik argue that this process does little to prevent migration, but does have political value by shifting the responsibility from politics to data and algorithms.

Furthermore, in their article on biometric technology used for border control, Stenum (2017) criticises the purpose- and function creep of migration-related databases. Purpose creep is described as using data for a different purpose than it was collected for. The author illustrates this with two changes to the EURODAC database which broadened its purpose from checking if an asylum seeker has applied for asylum in a different member state. First, it was expanded to also be used for anti-terror and anti-crime purposes. Then it was expanded again to also aid the identification of illegally staying third-country nationals to facilitate their return process. Function creep is described as an expansion of the functions of a database. This is exemplified by the introduction of facial images to EURODAC and the extension of the retention period of the data stored in this database (Stenum, 2017).

Finally, Molnar (2019) criticises the use of new technologies for migration management for the possibility to infringe on multiple fundamental human rights. First, decisions made on the basis of data and algorithms can result in the restriction of life and liberty by, for example, incarceration in detention centres. Secondly, the author highlights the increased risk immigrants face when their rights to privacy are violated such as in a situation where oppressive regimes, from which refugees are fleeing, obtain the data of those refugees. Lastly, there is a risk to impinge on equality rights through algorithmic discrimination, which the authors

deem likely due to the opaque nature of immigration and refugee decision making as well as the "problematic track record that automated technologies have on race and gender" (Molnar, 2019).

1.2 Frontex and public procurement

The organisation charged with border management across the EU is the European Border and Coast Guard Agency, also known as Frontex. Frontex was established in 2004 and was initially tasked with improving the coordination of border control at the external borders of EU member states (Council of European Union, 2004). Since 2015 however, Frontex has been growing in budget from €143 million in 2015 to €322 million in 2020, as well as staff size from 402 employees in 2016 to 1000 staff members in 2020. Frontex's mandate has also been expanded to cooperate with the member states in "implementing an integrated border management approach" (European Commission, 2015). Frontex has been criticized for lacking transparency, sometimes actively controlling the availability of information. This is illustrated by Frontex's frequent denial of freedom of information requests (FOI), the demanding of legal fees from activists submitting FOIs and requiring authorisation from Frontex itself before publishing documents obtained under FOIs (Karamanidou and Kasparek, 2020). This lack of transparency mirrors a general opacity of technology used at borders, where the closed source nature and extensive use of third party application programming interfaces (API) prevent easy insights into the inner workings of software used for migration purposes (Aradau et al., 2019).

There is, however, some information that Frontex is compelled to publish through EU regulation, that being public procurement documents. Whenever Frontex needs a good or service it can contract a third party via public procurement. Frontex does this by publishing a call for tender (CFT) which details the requirements of the contract subject, the budget of the contract, the contract procedure among other information. Prospective contractors can then bid on the contract and once Frontex chooses a company it has to publish a contract award notice (CAN), listing to whom the contract was awarded, when it was awarded, etc. For any contract with a value of over 139.000 Euro documents such as the call for tender and contract award notice need to be published (European Parliament and Council of the European Union, 2018).

1.3 Computational social science

These documents can be numerous and large in size, which is a lot of information to manually analyse. One approach is to use computers to assist this process. Using computational methods to assist social science research is what Lazer et al. (2009) argue for in their article titled "Computational social science". Grimmer and Stewart (2013) also espouse the use of automated text analysis for a political texts, while emphasising the role of manual validation.

One computational method that is used for social science is topic modelling. Topic modelling uses unsupervised machine learning models to automatically extract topics from a large corpus of documents. Each document can contain multiple topics and each topic is defined by a ranked list of words from the corpus. A popular model used in this field is Latent Dirichlet Allocation (LDA). LDA has been used to analyse journalistic texts (Jacobi et al., 2016), social network posts (Sokolova et al., 2016) and political texts (Zirn and Stuckenschmidt, 2014).

In brief, migration in the EU is becoming increasingly datafied and this process comes with a host of issues. However, the agency tasked with the management of the European borders, Frontex, is opaque. One way of acquiring insight into Frontex's technology strategy is by looking at publicly available procurement information, but this comes in the form of a large number of documents, making manual analysis infeasible. These documents can be analysed using topic modelling methods such as LDA. Most of the literature on datafication of borders refers to journalistic texts or case studies when illustrating the datafication in the EU. Thus, This research will attempt to give an empirical account of the datafication in the EU using topic modelling methods on a dataset of public procurement documents published by Frontex.

The second chapter of this paper will describe other research relevant to the methods used in this paper. The following chapter then describes how the documents are acquired and how data is extracted from these documents to create the corpora used for topic modelling. Next, the chapter titled method will go into detail on how these texts are processed and LDA is used to extract topics. The results of this process will be presented in the next chapter. Finally, in the last chapter, these results will be interpreted and limitations of this research are noted.

Chapter 2

Related work

2.1 Automated text analysis

Grimmer and Stewart (2013) argue that while there is a lot of value in analysing political texts, this is a difficult task due to the sheer volume of data often available. They maintain that automated text analysis can allow for processing large amounts of textual information without access to many resources. The authors do this by giving an overview of many different methods. They also introduce four principles of automated text analysis used to guide researchers in applying these methods. The first principle is "All quantitative models of language are wrong—but some are useful". This principle is based on the complexity of language which makes it almost certain that models will be wrong. This doesn't mean that models can't give useful insights, rather, the authors caution to know the limits of the models used. The second principle is "Quantitative Methods Augment Humans, Not Replace Them". Here, the authors argue that the methods will not eliminate the need for expert knowledge and analysis, but can enhance this process. In the third principle "There Is No Globally Best Method for Automated Text Analysis", Grimmer and Stewart contend that different datasets and research questions call for different text analysis methods. The final principle is "Validate, Validate, Validate". In this principle, the authors emphasize the need for validation and caution against simply using the result of any method without verifying it. Grimmer and Stewart also go into detail on how topic modelling methods can be validated. They note that topics need to be labelled and recommend doing so by, for example, reading a sample of documents associated with a given topic.

2.2 Latent Dirichlet Allocation

As mentioned in the introduction, topic modelling is a type of statistical modelling where the goal is to find topics in a large number of documents without specifying these topics beforehand. One popular topic model is Latent Dirichlet Allocation (LDA). LDA is based on the assumption that a corpus can be generated from two distributions: a distribution of words over topics and a distribution of topics over documents. That is, each topic is a distribution of words and each document is a distribution of topics. The parameters of this model are the counts of all words appearing in the corpus, the number of topics to be found (k), the prior of the document topic distribution (α) and the prior of the topic word distribution (β). The first output of a trained and fitted LDA model is an $n \times k$ matrix, where n is the number of documents and k is the number of topics. This matrix describes the distribution of topics over documents. For example, the first column of the sixth row describes the probability of the first topic generating from the sixth document. The second output of this model is a $k \times m$ matrix where k is the number of topics and m is the number of words in the vocabulary. This matrix describes the distribution of words over topics. In this matrix, the first column of the sixth row represents the probability of the first word generating from the sixth topic (Blei et al., 2003).

2.3 LDA evaluation metrics

LDA models are generally evaluated with likelihood-based metrics such as log-likelihood and perplexity. There is however some research that questions the utility of these metrics. Chang et al. (2009) developed two tasks to measure the coherence of a topic model and had these tasks executed by humans via a work crowdsourcing service for multiple models. These tasks were word intrusion and topic intrusion. For word intrusion, the top five words of a topic and with a random sixth word inserted is presented to the worker from which the worker must try to pick the intruder word. The idea behind this task is that the worker will pick the intruder word if the other five words are indeed coherent. If they are not coherent, the worker will have a good chance to pick one of the five words instead. The second task is topic intrusion which is very similar to word intrusion, but in this task, the worker is shown the title and a snippet from a document, the top three topics associated with that document and one intruder topic. The work must again try to find the intruder. The results of the study show that likelihood-based metrics such as perplexity don't always correlate positively with coherent topics as judged by humans (Chang et al., 2009).

Since the release of this article, multiple alternative coherence measures have

been suggested. One measure described by Mimno et al. (2011) is often referred to as the Umass coherence. In this research, the authors let two domain experts evaluate 148 topics created from a corpus of 300000 National Institutes of Health paper abstracts. The experts not only evaluated the topics as good, intermediate or bad but also tagged them with reasons for why they were bad. Examples of such reasons are that the topics contained seemingly random words or a few intruded words. Mimno et al. then developed a coherence metric to account for the specific reasons why topics were marked as bad. This metric is based on the co-occurrence of words within documents. The authors then compared the results of this metric with the expert evaluation as well as a replication of Blei et al.'s word intrusion experiment. They concluded that "It is possible to improve the coherence score of topics... all without requiring semi-supervised data or additional reference corpora." (Mimno et al., 2011).

2.4 Topic visualisation

Visualising topics for LDA has traditionally been done by presenting a list of words ranked by their probability to generate from that topic. In their paper Sievert and Shirley (2014) develop a novel visualisation based on dimensionality reduction. The visualisation is divided into two panes. In the left pane, the topics are represented as circles in a 2D grid. The distance between de topics represents how distinct the topics are compared to each other. The circles' size shows how large a portion of the corpus the topics encompass. The user can select a topic with their mouse and the right pane will show the list of ranked words that make up that topic, with the most important words on top. Sievert and Shirley also introduce a new metric to rank these words: relevancy. Relevancy is the weighted combination of the probability of a word and its lift, with the weight being a tuning parameter λ that can have a value between 0 and 1. A word's lift is defined as the ratio between a word's probability in a topic and its marginal probability in the whole corpus. Setting λ to 1 results in the traditional ranking while setting it to 0 results in ranking based solely on a word's lift. In other words, setting λ to 1 will result in the top words of the topic being words that more uniquely identify that topic. Sievert and Shirley also conducted a user study in which subjects were given topics represented by ranked word lists with different values of λ and had to choose which topic label the topic belonged to. This was then compared to the ground truth of those topics. This resulted in an optimal value for λ of about 0.6 (Sievert and Shirley, 2014).

Chapter 3

Dataset

The data was collected from two data sources from which two datasets were created: a contract award notice (CAN) dataset and a call for tender (CFT) dataset. From the latter, two variants were created, a full-text variant and a selected-section variant.

3.1 Contract award notices

Contract awards notices are documents that are published at the end of the procurement process. These documents contain information about the contract itself and to which organisation it was awarded as well as the award procedure. The documents are in XML format and are semi-structured. They contain mostly consistent fields, but the structure has slightly changed multiple times over the years. The fields are also filled in manually by the document author, so there are sometimes different spellings and interpretations of some fields or values. Which fields a CAN document contains differs depending on the contract and structure version. Each CAN generally contains the following sections:

- "Contracting authority" which contains information such as the official name, postal address and e-mail of the organisation issuing the CAN.
- "Object" contains information on what is to be delivered by the contractor such as a short description, a title, the total value of the contract and place of performance.
- "Procedure" which contains field such as "Type of the procedure", "Award criteria" and "Administrative information"

- "Award of contract" containing information like the name and address of the contractor, the date of conclusion of the contract and the value of the contract.

- "Complementary information". This section lists information about subcontracting, the date of dispatch of the CAN as well as information on which body will review the contract and provide mediation procedures.

The documents were obtained using the "expert search" feature on the TED website (*TED Tenders Electronic Daily*, n.d.) using the following query: "HA=[AG] AND AU=[*Border* OR *Frontex*] AND TD=[7]". The resulting 165 documents range from 2014 through 2021. Data is extracted using the python built-in XML module. This is done by traversing the XML tree using the "Element.find()" function with queries of the field names to be extracted. Some of the contracts are split into multiple lots or are awarded multiple times. In these cases, the CAN contains multiple award sections. Each of these award sections is treated as a single data point. The following information was extracted for each award: the contract ID, the contract object title, the type of contract, a short description of the contract object, the total contract value, the place of performance, the contractor and the notice dispatch date. For the LDA described in the methods section, only the short description was used as a corpus. The notice dispatch date was used for the creation of the timeline plots described in the methods section.

Seven documents were discarded because some elementary XML nodes could not be found. In total 287 data points were extracted. Figure 3.1 shows the distribution of these data points over the years.Using simple splitting on spaces, the short descriptions have a size of 14989 tokens in total. For an overview of the corpus sizes, see table 4.1.

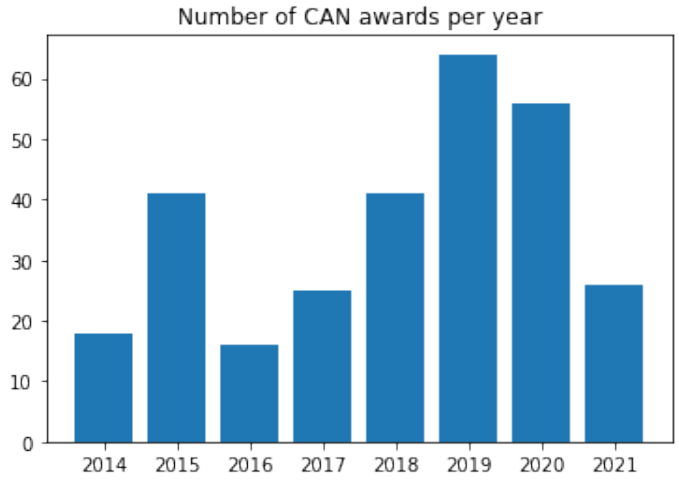


Figure 3.1: Distribution of contract award notices per year

3.2 Call for tender documents

Call for tender documents are published at the start of the procurement process and describe the contract and its object in detail as well as describing changes to the contract and contain important forms. Each CFT consists of multiple documents. We are interested in the document describing the tender in detail. These documents generally contain a detailed description of the contract object and its requirements as well as a description of the contract and the award procedure. They are generally referred to as Annex 2 or Terms of reference, but this is not always the case. The documents are somewhat structured: they are made up of multiple sections and many documents contain the same sections. However, many documents don't contain the same sections or have slightly different section names. These documents are in PDF format.

The documents were retrieved in the following way. First, the results page of a search on the "etendering" section of the TED website (*eTendering*, n.d.) was scraped. The search query was left empty and the "Contracting authority" field was set to "European Border and Coast Guard Agency". For each row of the results table, the hyperlink to the CFT page was extracted. From this hyperlink, the "cftid" was derived. This id was then inserted into the following hyperlink in place of the `[CFTID]`: `https://etendering.ted.europa.eu/document/archive-download.html?cftId=[CFTID]&lngIso=en`. This link was then used to download a ZIP file containing all documents related to a given CFT. From each ZIP file, a document containing some form of Annex 2 in the file name was then extracted. If

there was no document with Annex 2 in the file name, Annex 1 was used instead. In total, 126 documents were retrieved this way. Six documents were discarded because they either weren't the right document or they were scanned versions of documents which made it infeasible to extract the text.

PDF's, unlike other text file formats which encode text in plain text, are made up of drawing commands for drawing text, images or other graphics. This makes extracting text non-trivial and prone to errors. The text extracting was done using the python module "pdfminer.six". This library was chosen because it produced the best results for these documents. While the results were quite good, the output still contained some out of order text. It also included some extraneous whitespace such as spaces and newlines in seemingly random locations. For each document, the "Start date" was also extracted from each row in the table obtained by scraping the search results page for the initial links of the ZIP files. Both this date and the output from the text extraction were then saved into the full-text variant of the CFT dataset. In figure 3.2 the number of extracted documents per year is shown. The size of all full-texts combined is 770607 tokens when using simple splitting on spaces and ignoring tokens made up solely of whitespace.

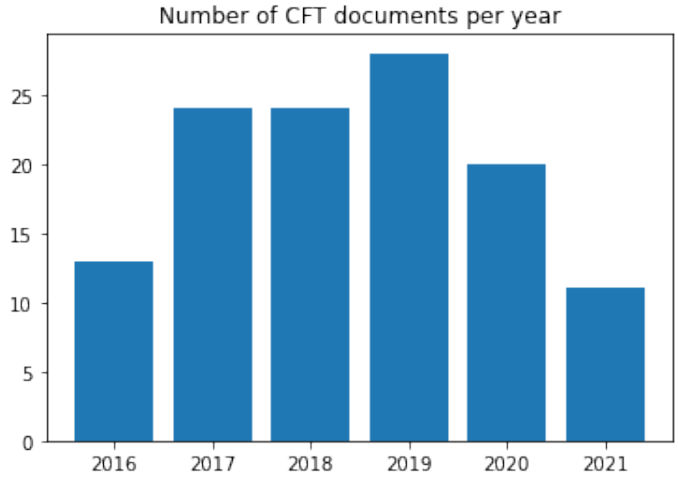


Figure 3.2: Distribution of call for tender documents per year

A second version of this dataset was also made by extracting only certain sections of each document. The intuition is to only extract information that is relevant to the object of the contract and leave behind information that concerns contract procedures, descriptions of Frontex itself and appendices such as forms. Since PDFs don't contain easily processable data on the structure of the document, the following method was chosen to extract the sections. For each document, an

ordered list was created manually containing the titles of all the sections in the document. The location of each section title was then found using the regex query `(\d|I|V)\.?\s+?[WORD]\s+?\n`. A given section was then defined as all the text between the location of the section title and the location of the next section title or the end of the document if the section title is the last title in the list. For each document, a list of sections to be extracted was then manually created. Some examples of extracted section titles are: "General information", "Specific/Technical Requirements", "Scope", "Background", "Current and target situation", "Quality requirements". Some examples of section titles that were left behind are: "Implementation of FWC", "Contract implementation", "Administrative Information", "Award of the contract", "Appendices", "Description of Frontex", "Table of Contents". The text of all selected sections was then concatenated and added as the corpus for this CFT in the selected-sections variant of the dataset. The size of this corpus when using the same simple tokenization method as the one used for the full-text variant is 517529 tokens.

Chapter 4

Method

The datasets created in the previous sections were used to train multiple LDA models. For each of the three corpora, the optimal hyperparameters were found. Then a model was trained with those parameters and the topics were labelled. The document-topic distributions and dates associated with a document were then used to create timeline plots for each topic.

4.1 CAN dataset

First, all documents with invalid or missing short description entries are removed. Then, the short description of each document was tokenized using the "spacy" python library to remove punctuation and stopwords. The resulting corpus contains 9683 tokens. See table 4.1 for an overview of the corpus sizes before and after processing. The "CountVectorizer" function from the "sklearn" python library was then used to create a word count matrix of shape $n \times m$. Here n is the number of documents and m is the number of unique words that occur in all texts combined. In order to find the right number of topics (k), an LDA model is trained multiple times with the following parameters: the word count matrix, an α and β of $1/k$ (Rieger et al., 2020) and each k in the range of 2 through 50. The resulting log-likelihood-, perplexity- and coherence-scores are then graphed with the range of k values on the x-axis. The three optimal values for k are taken to be the ones on which the optimums of the three curves agree.

	CAN corpus	CFT full-text corpus	CFT selected-sections corpus
Number of tokens before processing	14989	770607	517529
Number of tokens after processing	9683	453539	297220

Table 4.1: Sizes of the corpora before and after tokenization and filtering

Three LDA models are then trained and fitted using the three chosen values of k. The models are inspected manually using the "pyldavis" python module with the λ set to 0.6 as a starting point. For each topic, a label is chosen by examining the ranked word list representing the topic. The value of λ is manually adjusted when necessary to get a good understanding of the topics found. The labels of a given topic are then fine-tuned by looking at a sample of documents for which this topic ranked the highest. The three models with the labelled topics are then shown to a domain expert from an EU funded research group on datafication of borders. Together the final model is chosen and the labels are tweaked. The following quote is an example of the kind of feedback received: "I've had a look at the keywords and K9 looks like the best summary to me – it is both comprehensive but specific enough. Lower than this, I think it would miss some key topics. Higher than this, it seems to create topics that don't seem as logical to me/or have quite mixed words." Finally, for the CAN dataset, a model with 9 topics was chosen.

For the timeline analysis, all documents were grouped per year of publication. For each group, the mean along the first axis is taken of the $k \times n$ matrix representing the topic distributions of all documents where n is the number of documents. That is, for each topic the mean is taken of that topic's contribution to all documents in the year group. This results in a $k \times 1$ representing the topic distribution of that year. Using these vectors k plots are made, one for each topic, showing the change in the contribution of that topic over the years.

4.2 CFT dataset

For the full-text variant of the CFT dataset, all documents with invalid or missing entries are removed. The text was then also tokenized using the spacy python module. The text extracted from the documents contains many nonsense tokens owing to tables, figures or section headings in the original PDF. Because of this, all tokens made up solely of whitespace or containing digits are filtered from the corpus resulting in a corpus containing 453539 tokens. Then the optimal k, topics and topic labels are found with the same method as the one used for the CAN dataset. The topic timeline plots are also created in the same manner as for the CAN dataset. This is then also done for the selected-section variant of the dataset. The selected-sections corpus had a size of 297220 tokens after processing. For the CFT full-text dataset a k of 8 was chosen and for the selected-sections variant a k of 9 was chosen.

Chapter 5

Results

In this section, the topics of the resulting LDA models for each dataset are presented along with their respective timeline plots.

5.1 CAN dataset

The topic model trained on the CAN dataset produced nine topics, which are listed in table 5.1 (see appendix A for the extended word lists). For each word, the top 15 words are listed in ranked order. The top 30 words for the topic information systems includes words such as eurosur, system, communication, network and graphic. This topic refers to documents about systems to store and share information. Local operational support contains words such as fuel, car, operating, serbia equip and fruit. This topic is mostly found in documents concerning support Frontex offers to border control agencies of member states. Organised events involves words such as conference, event, lot, participant and invitation. Border checkpoint has words such as office, mobile, document, scanner, fingerprint and language. This topic is likely about border checkpoints themselves, where crossing migrants are checked and registered. The topic labelled contract consists of words like service, provision, contract, procedure, framework and award. This topic covers mostly contractual information. The software development topic contains words like software, development, maintenance, source, system, technology and refactoring. Surveillance is comprised of words such as surveillance, area, equipment, arial, maritime, border, event, report and flight. This topic concerns surveillance, be that from the air, on the ground or on the seas. It also seems to reference targets of surveillance such as areas or events. The topic labelled ground support/equipment contains words like situation, crisis, greece, vehicle, emergency, site and road. This topic seems similar to local operational support but appears to deal more with exceptional situations, rather than general operational activities. The final topic, which

information systems	**local operational support**	**organised events**	**border checkpoint**	**contract**	**software development**	**surveillance**	**ground support /equipment**	**clothing /uniforms**
eurosur	fuel	organisation	office	service	software	surveillance	situation	unisex
system	provision	conference	container	use	development	frontex	ketrzyn	winter
communication	card	lot	associate	provision	maintenance	level	koszalin	shirt
sensor	car	event	polish	contract	service	area	crisis	0
network	operating	poland	renting	work	contract	define	need	jacket
graphic	serbia	100	mobile	procedure	shall	provide	greece	guard
camera	insure	participant	toilet	annex	domain	interest	vehicle	risk
control	man	invitation	delivery	competition	source	equipment	frontex	cap
print	properly	outside	document	frontex	support	object	operational	sports
material	associate	divide	survival	framework	consider	maritime	emergency	polo
design	equip	responsible	scanner	reference	system	aerial	leasing	trousers
internal	return	range	language	warsaw	etc	border	activity	suit
base	research	scope	law	seat	solution	designate	site	socks
template	study	tender	skill	ceiling	understand	eu	staff	jogging
ms	fruit	ii	device	fwc	price	land	professional	summer

Table 5.1: Topics from the CAN dataset model. Top 15 words ranked with $\lambda = 0.6$.

is labelled clothing/uniforms contains words such as unisex, winter, shirt, guard, polo and trousers.

Figure 5.1 shows the timeline plots of the CAN dataset topics. Each plot shows the contribution of the given topic to the documents of that year. For example, the top left plot shows that in 2014 the topic information systems topic had a relatively little probability to occur in the documents of that year, while its proportion in 2017 was relatively large, growing to around 0.2 probability. In the following years, the topic became less important towards 2020, after which it saw an uptick in 2021. Local operational support started relatively unimportant until 2017 after which it grew to a probability of 0.1. The organised events topic has a large spike to about 0.3 in 2015, after which its importance decreased. Although it has a small uptick in 2018. The topic labelled border checkpoint started with a dip in 2015 but then slowly grew to a peak of 0.3 probability in 2020, but declined sharply in 2021. The contract topic has a very similar shape the only difference being that it started a lot higher with a probability of around 0.3 in 2014. Software development also started with a dip from around 0.2 probability in 2014 to 0.1 in 2015. After this, it quickly rises to around 0.4 in 2017. Then there is a sudden and large drop in 2018 after which it rises over the next years back to 0.4 in 2021. The surveillance topic seems to follow a roughly inverted shape, starting with a small rise from around 0.25 to 0.3 probability followed by a drop to almost zero in 2017. After this, there is a quick increase towards about 0.25 in 2018 and then a slow descent to about 0.75 in 2020 and 2021. The bottom middle plot shows that the ground support/equipment topic stays at around 0.1 probability throughout all years except for a spike to around 0.2 in 2016. The last plot also

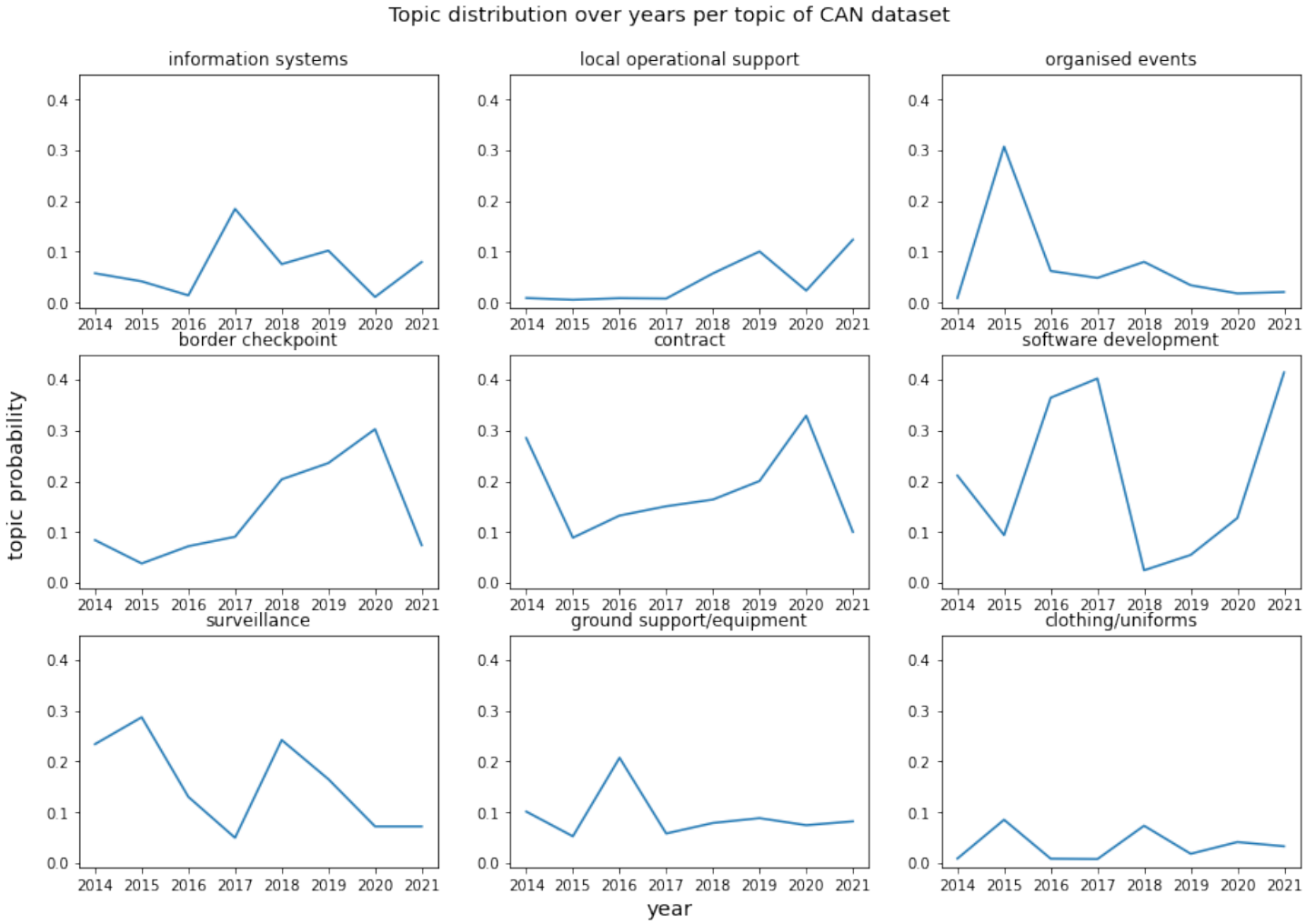


Figure 5.1: Timeline plots for each topic from the CAN dataset model

shows the clothing/uniforms topic to hover around a relatively low probability, with the exception of two spikes in 2015 and 2018 which reach a probability of 0.1.

5.2 CFT full text-dataset

The eight topics found in the CFT full text dataset are represented in table 5.2. Some topics are similar to the CAN dataset topics. These topics are contracting, software development, information systems, surveillance and local operational support. There are, however, some differences in the words representing these topics. The words in the CFT dataset seem to be more specific. For example, information

equipment /interior	**contracting**	**software development**	**information systems**	**surveillance**	**training**	**hardware**	**local operational support**
pistol	frontex	frontex	system	service	cme	hp	container
furniture	shall	shall	user	frontex	crisis	fio	office
shall	vehicle	software	hr	contract	exercise	bladesystem	facility
ammunition	warsaw	document	allow	contractor	course	factory	toilet
magazine	europejski	project	te	flight	textbook	fibre	mobile
paint	eu	development	ms	shall	teacher	kit	deployable
mm	pl	solution	se	provide	semester	gb	month
mount	contractor	contractor	sac	mission	participant	hba	space
trigger	www	task	ahr	specific	language	adptr	cleaning
holster	europa	information	deployment	request	post	fc	year
clean	tel	use	opera	hour	management	brocade	associate
delivery	european	feature	resource	datum	listen	svc	furniture
malfunction	fax	management	sgo	tender	group	ultrium	montenegro
supply	poland	level	type	time	headquarter	bl	service
frontex	coast	support	need	surveillance	frontex	blade	equipment

Table 5.2: Topics from the CFT full-text dataset model. Top 15 words ranked with $\lambda = 0.6$.

systems in the CAN dataset contains words such as communication, system and network while the same topic in the CFT full text dataset contains words like user, allow, resource, browse and activity. The contracting topic in the full text dataset also seems to contain more words representing, for example, contact details. One of the new topics is equipment/interior and consists of words like pistol, furniture, paint, ammunition, flashlight, malfunction and floor. This seems to concern both combat-related equipment as well as interiors. Another topic that is not in the CAN dataset is labelled training and contains words such as crisis, exercises, textbook, teacher, feedback and exam. This topic seems to pertain to following or providing certain courses. The hardware topic consists of words like hp, factory, fibre, blade, pcie, microsd and gpu. Most of these terms are related to computer hardware.

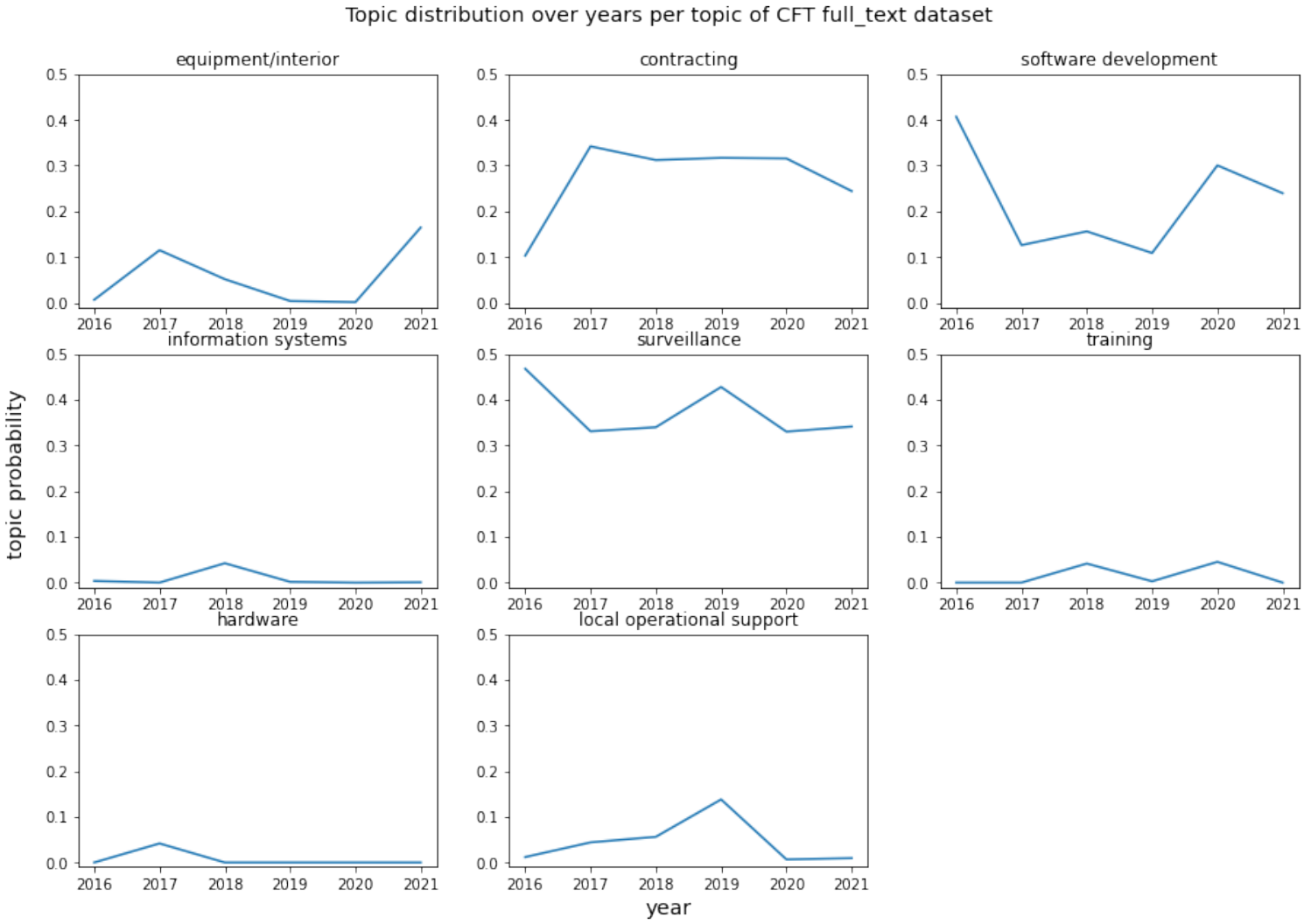


Figure 5.2: Timeline plots for each topic from the CFT full-text dataset model

The timeline plots of the CFT full text dataset topics are shown in figure 5.2. It is important to note the difference in the timeline between the CAN and CFT datasets when comparing the timeline plots, the CAN dataset goes back to 2014 while the CFT datasets start in 2016. The equipment/interior topic starts with a very low probability, rising to about 0.1 in 2017 and dipping down again in the years 2019 and 2020. It ends by rising to about 0.15 in 2021. The contracting topic starts at around 0.1 probability and rises to around 0.35 in 2017. It then stays on this level until it decreases to about 0.25 in 2021. This is a slightly different shape from the contracting topic in the CAN dataset which has a slower rise and a peak around 2020. Software development starts at a probability of about 0.4 but

fresh food	**border checkpoint**	**branding**	**software**	**organised events**	**surveillance**	**weather /maps**	**crisis management**	**local operational support**
skin	feature	pantone	frontex	canteen	flight	service	cme	shall
fruit	chip	blue	contractor	room	mission	shall	crisis	vehicle
ripe	rfid	reflex	shall	furniture	surveillance	map	exercise	frontex
bright	module	colour	contract	service	deployment	datum	post	pistol
insect	scan	logo	service	operator	lot	frontex	preparedness	warranty
vegetable	readable	foil	management	floor	service	forecast	headquarter	tel
malformation	traveller	cardboard	document	office	interest	provide	rehearse	europejski
carrot	software	print	system	facility	contract	weather	session	pl
fresh	display	pcs	project	coffee	area	request	awareness	warsaw
mature	fingerprint	paper	work	catering	aircraft	travel	management	minimum
mould	read	grey	task	building	authority	time	feedback	poland
corrugation	document	matt	software	event	container	satellite	annual	delivery
decay	light	black	user	conference	datum	oceanographic	intersection	europa
rind	border	pocket	support	hotel	crew	agency	cmes	www
bruise	identification	inner	solution	equipment	coordination	ship	fledge	european

Table 5.3: Topics from the CFT selected-sections dataset model. Top 15 words ranked with $\lambda = 0.6$.

quickly falls to around 0.1 in 2017. It stays around that level until 2020 where it rises to around 0.3 and ends at about 0.25 in 2021. This is a similar shape to the topic in the CAN dataset which also starts high around 2016, then dips and regains its importance towards 202021. Information systems stays at 0 probability for all years except 2018, where it has a peak of about 0.05. The surveillance topic starts high at around 0.5 probability after which it decreases slightly to about 0.3 where it stays for the remaining years with the exception of 2019 where it has a peak of around 0.4. This topic also has a very similar shape to its CAN counterpart, albeit with a higher base probability. The topic labelled training stays around 0 probability for all years except for two small peaks to around 0.05 in 2018 and 2020. The hardware topic also stays around 0 probability with a peak of about 0.05 in 2017. The local operational support topic starts around 0 probability in 2016 and rises to around 2019 over the years after which it drops down again to around 0 for the remaining years.

5.3 CFT selected sections-dataset

The topics found for the selected sections variant of the CFT dataset can be seen in table 5.3. The border checkpoint and organised events topic are similar to the topics with the same label in the CAN dataset. Like the full text dataset, these topics seem to contain more specific words compared to CAN dataset. The software, surveillance and local operational support topics can also be found in the

CAN and full text dataset. The first new topic, fresh fruit, contains words like fruit, ripe, mould and vegetable. Branding consists of words like blue, colour, logo, print and pocket. The topic labelled weather/maps consists of words like map, forecast, weather and oceanographic. This topic seems to be about information related to travel by boat such as weather conditions. Crisis management contains words such as crisis, exercise, awareness, preparedness and management. This seems to pertain to training staff to deal with certain emergency situations. There is quite some overlap with the training topic from the full text dataset.

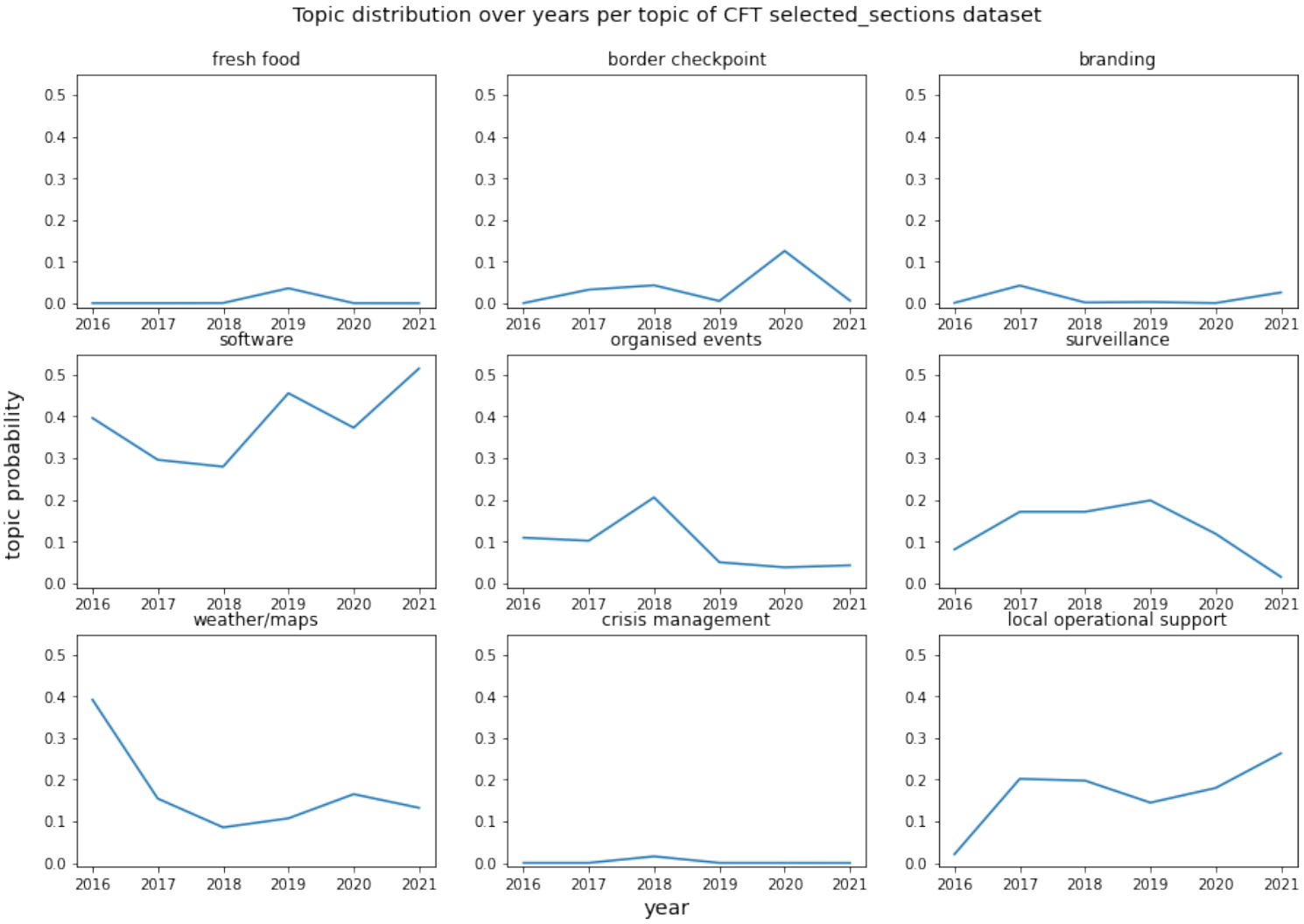


Figure 5.3: Timeline plots for each topic from the CFT selected-sections dataset model

The timeline plots for the selected sections dataset are shown in figure 5.3.The fresh food topic is relatively unimportant except for a small peak in 2019. Border checkpoint has a small bump reaching around 0.05 over the years 2017 and 2018 and a larger peak to around 0.1 in 2020. Except for the dip in 2019, this is a similar shape to the CAN dataset counterpart of this topic, albeit at an overall lower probability. The topic labelled branding starts with a small peak to 0.05 in 2017 and ends with a small rise towards 2021. The software topic stays important throughout the years, never going below about 0.3 probability. Rises from about 0.4 in 2016 to about 0.5 in 2021, but has two dips in 2018 and 2020. Compared to the same topic in the CAN dataset, this topic seems to have less variance but exhibits the same general rising trend. Compared to the full text dataset, this topic starts out very similar but seems to have an almost inverted shape from 2018 onwards. Organised events stays level at around 0.1 probability for the years 2016 and 2017. It then grows to a peak of 0.2 but quickly falls to around 0.05 for the remaining years. The shape of this development is almost identical to the topic labelled organised events from the CAN dataset. The surveillance topic takes the shape of a large bump starting around 0.1 probability in 2016, rising to 0.2 in 2019 and then dropping down to around 0 in 2021. The topic labelled weather/maps starts high at around 0.4 probability and slowly decreases to 0.1 in 2018 after which it rises slightly to 0.15 for the remaining years. Crisis management has a small peak in 2018. Finally, the local operational support topic starts around 0 probability in 2016 but rises to around 0.2 in 2018 and stays around that level for the remaining years. This is different from the CFT counterpart in shape, but similar to the CAN topic of the same name.

Chapter 6

Discussion and conclusion

6.1 Interpretation of results

Examining the differences between the results of the different models we find that generally many of the same topics were found in all models. These models did contain slightly different words with the words in the CFT topics being more specific. This makes sense because the CAN dataset contained only short descriptions of the contract object so we would expect these to be a bit higher level than the more detailed CFT contents. This means that using this method can be used on corpora of different levels of abstractions depending on the kind of information that is relevant. Comparing the full-text to the selected-sections we also see that creating the second variant had the intended effect. We were able to find more specific topics in the selected-sections model while not finding a distinct contracting topic, which we did find in the full-text model. The words generally associated with contracting also didn't intrude the other topics in the selected-sections model to compensate for the lack of explicit contracting topic. Thus one can improve the topic models by filtering larger pieces of texts known to contain irrelevant information from the corpus.

Inspecting the topics and their resulting timelines, local operational support appears to be a stable topic, occurring in all three models. This is expected since this is one of Frontex's main tasks according to their expanded mandate. As such, we also see the importance of this topic generally increasing over the years. Topics such as branding, clothing/uniforms and training generally don't vary a lot and seem to become relevant periodically. This is what is to be expected as while the size of those contracts might expand, the proportion compared to other topics would not. These are also goods/services that would be renewed every couple of years. This gives us confidence that the method produces topics that are distinct, coherent and represent Frontex's contracting actions. Surveillance and software

development appear to be major topics in the contracting Frontex does. They are distinct topics that appear in every model and usually have a large contribution to the documents compared to other topics. Across the years these topics do tend to have quite some variance. However, broadly speaking, whenever one of the topics seems to decrease in probability the other seems to increase. This might be explained by a cadence of acquiring methods to obtain data (surveillance) followed by acquiring methods to process this information (software development). This supports the idea of marketization of the datafication of borders. A notable absence in the topics and their associated words are human rights or transparency, which indicates that this is not one of the higher priority subjects when procuring goods and services from third parties. Lastly, the presence of words and topics relating to crises, weapons and equipment support the crisis framing of migration.

6.2 Limitations

There are some limitations of this research. First of all, there are some limitations of the datasets used. One of those limitations is the size of the short descriptions in the CAN dataset which can limit the number of topics to be found as well as their specificity since LDA is based on the assumption that each document is based on a mixture of topics. Second, both the CAN and CFT datasets have relatively few documents. This means that certain topics might only occur in one or two documents in the corpus, which might make it harder to find those topics. Next, the documents range from a limited timeline compared to Frontex's overall existence. It might be more useful to see the progression of certain topics over a longer period. Another limitation is the fact that, as per figure 3.1 and figure 3.2, the distribution of documents per year is not uniform. There are, for example, relatively few documents for the year 2016. This might cause the contribution of certain topics to a given year to vary more since a single document has more effect on the proportion in a year with fewer documents than a document in a year with more documents. Finally, due to the CFT being published at the start of the procurement procedure and the CAN being published at the end, the timeline plots of CANs would be shifted compared to the CFT timeline plots. This shift would also be different for each contract depending on the length of the procurement procedure.

Second of all, some limitations of the method should be noted, one being that we do not know what goods and services are contracted and what is done internally by Frontex's. This might give a skewed perspective on the importance of certain topics. Related to this is the fact that only contracts over a certain threshold need to be published. This means that many smaller but related contracts will be missed by using this method.

6.3 Conclusion and future research

We conclude that this method succeeds in finding meaningful topics in procurement documents that say something about Frontex contracting actions. This research also gives an empirical account of datafication, while also giving support to some of the theories around datafication of borders found in the social science literature.

Possible future research could try to combine these topic distributions with other available data such as the budget and the awarded contractor. This could then be used to, for example, use the contract value to assign a monetary value to the topics or identify contractors associated with certain topics. Another possible direction is extending the dataset to include procurement data from other sources or organisations such as the eu-LISA, the EU agency for IT.

Bibliography

Aradau, C., Blanke, T. and Greenway, G. (2019), 'Acts of digital parasitism: Hacking, humanitarian apps and platformisation', *new media & society* **21**(11-12), 2548–2565.

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003), 'Latent dirichlet allocation', *the Journal of machine Learning research* **3**, 993–1022.

Broeders, D. and Dijstelbloem, H. (2015), The datafication of mobility and migration management: The mediating state and its consequences, *in* 'Digitizing identities', Routledge, pp. 242–260.

Chang, J., Boyd-Graber, J., Wang, C., Gerrish, S. and Blei, D. M. (2009), Reading tea leaves: How humans interpret topic models, *in* 'Neural information processing systems', Vol. 22, Citeseer, pp. 288–296.

Council of European Union (2004), 'Council regulation (EU) no 2007/2004'. `https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX\%3A32004R2007`.

Csernatoni, R. (2018), 'Constructing the eu's high-tech borders: Frontex and dual-use drones for border management', *European Security* **27**(2), 175–200.

eTendering (n.d.), `https://etendering.ted.europa.eu/`. Accessed: 2021-06-25.

European Commission (2015), 'European agenda on migration: Securing europe's external borders'. `https://ec.europa.eu/commission/presscorner/detail/en/MEMO_15_6332`.

European Parliament and Council of the European Union (2018), 'Eu regulation no 2018/1046'. `https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=uriserv:OJ.L_.2018.193.01.0001.01.ENG`.

Grimmer, J. and Stewart, B. M. (2013), 'Text as data: The promise and pitfalls of automatic content analysis methods for political texts', *Political analysis* **21**(3), 267–297.

Jacobi, C., Van Atteveldt, W. and Welbers, K. (2016), 'Quantitative analysis of large amounts of journalistic texts using topic modelling', *Digital Journalism* **4**(1), 89–106.

Karamanidou, L. and Kasparek, B. (2020), 'Fundamental rights, accountability and transparency in european governance of migration: The case of the european border and coast guard agency frontex'.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M. et al. (2009), 'Social science. computational social science.', *Science (New York, NY)* **323**(5915), 721–723.

Metcalfe, P. and Dencik, L. (2019), 'The politics of big borders: Data (in) justice and the governance of refugees', *First Monday* .

Mimno, D., Wallach, H., Talley, E., Leenders, M. and McCallum, A. (2011), Optimizing semantic coherence in topic models, *in* 'Proceedings of the 2011 conference on empirical methods in natural language processing', pp. 262–272.

Molnar, P. (2019), 'Technology on the margins: Ai and global migration management from a human rights perspective', *Cambridge International Law Journal* **8**(2), 305–330.

Rieger, J., Rahnenführer, J. and Jentsch, C. (2020), Improving latent dirichlet allocation: On reliability of the novel method ldaprototype, *in* 'International Conference on Applications of Natural Language to Information Systems', Springer, pp. 118–125.

Sievert, C. and Shirley, K. (2014), Ldavis: A method for visualizing and interpreting topics, *in* 'Proceedings of the workshop on interactive language learning, visualization, and interfaces', pp. 63–70.

Sokolova, M., Huang, K., Matwin, S., Ramisch, J., Sazonova, V., Black, R., Orwa, C., Ochieng, S. and Sambuli, N. (2016), 'Topic modelling and event identification from twitter textual data', *arXiv preprint arXiv:1608.02519* .

Stenum, H. (2017), 'The body-border. governing irregular migration through biometric technology', *spheres: Journal for Digital Cultures* **4**, 1–16.

Taylor, L. and Meissner, F. (2020), 'A crisis of opportunity: Market-making, big data, and the consolidation of migration as risk', *Antipode* **52**(1), 270–290.

TED Tenders Electronic Daily (n.d.), `https://ted.europa.eu/`. Accessed: 2021-06-25.

Zirn, C. and Stuckenschmidt, H. (2014), 'Multidimensional topic analysis in political texts', *Data & Knowledge Engineering* **90**, 38–53.

Appendix A

Extended Topic tables

information systems	**local operational support**	**organised events**	**border checkpoint**	**contract**	**software development**	**surveillance**	**ground support /equipment**	**clothing /uniforms**
eurosur	fuel	organisation	office	service	software	surveillance	situation	unisex
system	provision	conference	container	use	development	frontex	ketrzyn	winter
communication	card	lot	associate	provision	maintenance	level	koszalin	shirt
sensor	car	event	polish	contract	service	area	crisis	0
network	operating	poland	renting	work	contract	define	need	jacket
graphic	serbia	100	mobile	procedure	shall	provide	greece	guard
camera	insure	participant	toilet	annex	domain	interest	vehicle	risk
control	man	invitation	delivery	competition	source	equipment	frontex	cap
print	properly	outside	document	frontex	support	object	operational	sports
material	associate	divide	survival	framework	consider	maritime	emergency	polo
design	equip	responsible	scanner	reference	system	aerial	leasing	trousers
internal	return	range	language	warsaw	etc	border	activity	suit
base	research	scope	law	seat	solution	designate	site	socks
template	study	tender	skill	ceiling	understand	eu	staff	jogging
ms	fruit	ii	device	fwc	price	land	professional	summer
product	macedonia	helpdesk	fingerprint	award	tuning	report	insurance	ed4bg
display	republic	venue	chair	point	meaning	specific	road	raising
powerpoint	bulgaria	service	capable	aircraft	mean	event	programme	21
layout	innovation	describe	service	union	technology	external	psychological	800
leaflet	charter	annex	course	initial	life	objective	attention	2015
word	cash	term	gift	negotiate	cycle	coordination	bari	budgetary
brochure	notice	oceanographic	portable	11	typical	contractor	immediate	eur
thermal	systematic	marine	speaking	repetition	broad	centre	independent	rap
node	maintain	atmospheric	foreigner	fr	refactoring	pre	break	article
border	carry	subject	semester	outset	engineering	frame	covid	134
include	fully	relate	everyday	journal	fix	availability	quarantine	procurement
publication	evaluation	storage	case	year	technical	tender	emotional	short
national	short	providing	reading	announce	assignment	flight	19	amendment
handheld	north	provision	listening	official	cover	enable	hotline	organise
cool	technological	condition	perform	similar	phase	course	exercise	softshell

Table A.1: Topics from the CAN dataset model. Top 30 words ranked with $\lambda = 0.6$.

equipment /interior	**contracting**	**software development**	**information systems**	**surveillance**	**training**	**hardware**	**local operational support**
pistol	frontex	frontex	system	service	cme	hp	container
furniture	shall	shall	user	frontex	crisis	fio	office
shall	vehicle	software	hr	contract	exercise	bladesystem	facility
ammunition	warsaw	document	allow	contractor	course	factory	toilet
magazine	europejski	project	te	flight	textbook	fibre	mobile
paint	eu	development	ms	shall	teacher	kit	deployable
mm	pl	solution	se	provide	semester	gb	month
mount	contractor	contractor	sac	mission	participant	hba	space
trigger	www	task	ahr	specific	language	adptr	cleaning
holster	europa	information	deployment	request	post	fc	year
clean	tel	use	opera	hour	management	brocade	associate
delivery	european	feature	resource	datum	listen	svc	furniture
malfunction	fax	management	sgo	tender	group	ultrium	montenegro
supply	poland	level	type	time	headquarter	bl	service
frontex	coast	support	need	surveillance	frontex	blade	equipment
flashlight	border	system	pru	information	preparedness	flexfabric	provision
slide	guard	requirement	select	offer	coordinator	microsd	door
accessory	agency	training	implement	lot	lesson	pcie	ukraine
warsaw	contract	contract	prepare	area	learn	gpu	floor
equivalent	delivery	border	proposal	event	shall	fan	instal
test	fwc	technology	opd	cost	feedback	numbers	cyprus
spire	order	work	budget	tenderer	rehearse	cto	kosovo
verification	equipment	quality	plan	include	session	em	moldova
cleaning	minimum	user	browse	authority	level	serial	fruit
shoot	provide	fwc	activity	require	awareness	dl	deployment
weapon	warranty	deliverable	modify	day	fledge	exp	specific
semi	requirement	application	filter	interest	exam	ltu	fresh
armourer	camera	datum	create	financial	framework	ctr	location
surface	deliver	want	request	agency	survival	xeon	installation
floor	technical	site	operational	centre	cmes	dmr	spain

Table A.2: Topics from the the CFT full-text dataset model. Top 30 words ranked with $\lambda = 0.6$.

fresh food	**border checkpoint**	**branding**	**software**	**organised events**	**surveillance**	**weather /maps**	**crisis management**	**local operational support**
skin	feature	pantone	frontex	canteen	flight	service	cme	shall
fruit	chip	blue	contractor	room	mission	shall	crisis	vehicle
ripe	rfid	reflex	shall	furniture	surveillance	map	exercise	frontex
bright	module	colour	contract	service	deployment	datum	post	pistol
insect	scan	logo	service	operator	lot	frontex	preparedness	warranty
vegetable	readable	foil	management	floor	service	forecast	headquarter	tel
malformation	traveller	cardboard	document	office	interest	provide	rehearse	europejski
carrot	software	print	system	facility	contract	weather	session	pl
fresh	display	pcs	project	coffee	area	request	awareness	warsaw
mature	fingerprint	paper	work	catering	aircraft	travel	management	minimum
mould	read	grey	task	building	authority	time	feedback	poland
corrugation	document	matt	software	event	container	satellite	annual	delivery
decay	light	black	user	conference	datum	oceanographic	intersection	europa
rind	border	pocket	support	hotel	crew	agency	cmes	www
bruise	identification	inner	solution	equipment	coordination	ship	fledge	european
ridge	automatic	clip	provide	cleaning	contractor	hp	lesson	road
flesh	information	pc	development	kitchenette	tender	parameter	participant	mm
free	mrz	imprint	use	clean	specific	border	outline	coast
orange	training	ring	require	participant	maritime	image	resume	requirement
texture	radio	lie	include	premise	tenderer	sea	making	contractor
thickness	comparison	stitch	request	meeting	hour	information	alia	deliver
disease	frequency	white	level	spire	centre	meteorological	interdependency	guard
peel	req	page	specific	lunch	aerial	format	reparation	eu
smell	ees	matching	perform	kitchen	contracting	priority	misunderstand	acceptance
spongy	nr	metal	security	frontex	airport	available	bc	fax
baton	accessory	prefolde	staff	agency	rpas	cloud	repatriation	border
plump	gate	rounded	business	storage	asset	requirement	mgmt	camera
overly	check	cover	plan	drink	designate	precipitation	biannual	thermal
firm	image	expandable	technical	staff	provide	imagery	reputational	transportation
bird	core	elastic	relate	delivery	payload	wind	taker	technical

Table A.3: Topics from the CFT selected-sections dataset model. Top 30 words ranked with $\lambda = 0.6$.