METHODOLOGY OUTLINE

PRESS 2022

4 Steps to a comprehensive classification

1. Reading the CRS data
2. Preparing the data
3. **A:** Title pattern matching

    3.1 Minority language treatment

4. **B:** Text mining of long descriptions

    4.1 Technical aspects of text mining

    4.2 Performance checks

1. Reading the CRS data

After downloading all .txt files for years 2006 - 2020 from

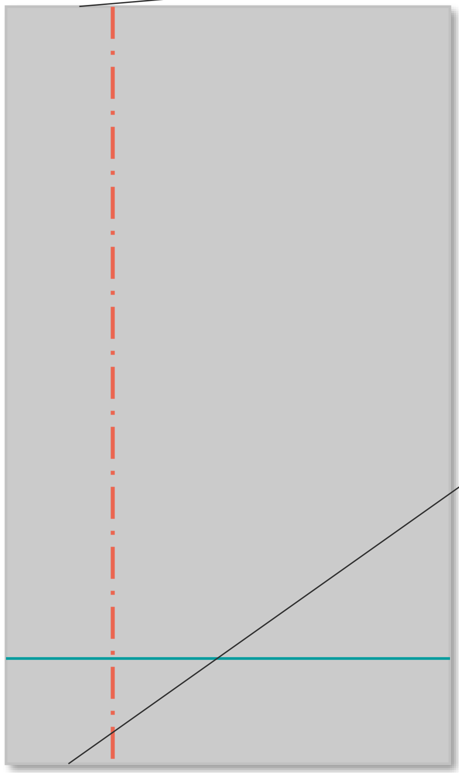
https://stats.oecd.org/DownloadFiles.aspx?DatasetCode=CRS1 ,

fully merged data set is stored as a .rds file.

2. Adding unique text identifiers

Unique text id’s are used in the process to reduce computation time since only unique titles and long descriptions are processed and later joined to full data set acc. to text id’s

Full CRS data



Drop French part of titles/descriptions for Canadian projects

Reduce full CRS data to subset of 16 necessary variables

Reduced CRS data

projecttitle longdescription | title_id text_id | descr2mine | Identifiers (scb, pop, gen_donor…) ….

1. Text cleaning | 2. Id creation | 3. descr2mine | 4. Creating identifiers

1. **Text cleaning**: numbers & punctuation removed, all lowercased
2. **Id creation:** unique title/description identifier through hashing of strings using xxHash (see https://github.com/Cyan4973/xxHash for details), outputs string of length 8
3. **descr2mine:** use only distinct long description (that differ from title) for text mining, compare description with title by counting number of alterations that would be needed to get to the title string (see Damerau-Levenshtein-Distance for details) -> if the string distance is lower than 10, this distinct long description is used for text mining
4. **Creating identifiers:** identifiers for statistics (scb = 16062, pop = 13010) and gender (gen_ppcode = 15170:15180, gen_maker, gen_donor for channel code = 41146 (UN Women), gen_sdg = sdg_focus contains 5)

3. Title pattern matching

Language detection: using Google's Compact Language Detector 2 (cld2), every project title's and long description's language is detected, NA usually comes from too short strings

Reduced CRS data

title_id	title_language	long_language	....
342ex	en	de	
ee455	en	en	
33x5d	sa	NA	
290r4	NA	fr	
3345t	es	es	
3qq56	nl	de	
44qw3	ol	fr	
4550e	NA	nl	
3369w	NA	en	
	:		
	:		

1. **Language filtering:** only combinations (en, es, fr, de NA) x (en, es, fr, de, NA), exclude (NA, NA)
2. **Drop duplicated title id's:** to analyze the same title only once

Data set for title pattern matching

title_id	title_language	long_language	....
342ex	en	de	
ee455	en	en	
290r4	NA	fr	
3345t	es	es	
3369w	NA	en	
	:		
	:		

statistics	gender	keywords
TRUE	FALSE	statistics
FALSE	FALSE	-
TRUE	TRUE	dhs, girl
	:	
	:	

Add detection filters to original reduced data set acc. to title id

Title pattern matching

1. **Clean and lemmatize keyword lists:** 4 keyword lists (statistics, gender, mining, stat. acronyms) are lemmatized (e.g. women's -> woman)
2. **Clean and lemmatize titles**
3. **Keyword detection:** for every title, 4 logical variables (one for each detection class) are set to TRUE if keyword is found in the lemmatized title; for acronyms original title is used
4. **Merging classes for final statistics and gender filter:** detection_statistics is set to TRUE **if** statistics is TRUE or stat. acronyms is TRUE or SCB is TRUE and mining is FALSE (exclude mining since "mining surveys" don't related to statistics), detection_gender is TRUE **if** gender was detected

3.1 Particularities for minority language

Minority languages are the three most frequent non-English languages: **French, Spanish, German**

The procedure is the same as on the previous slide except for the following adjustments:

- **Stemming instead of lemmatization:** for minority languages, the are currently no good lemmatization packages available, therefore stemming is used
- **Compound words treatment for German:** in German, many nouns are composed of two or more other nouns which is not accounted for by the stemming algorithm. Therefore, the nouns on the keyword list are detect also **within** compound words (e.g. “women initiative” -> “Haueninitiative” -> detected as gender because of “Frau”)

4. Text mining of long descriptions

Reduced CRS data

text_id	long_language	statistics	....
342ex	en	TRUE	
ee455	en	TRUE	
33x5d	sa	FALSE	
290r4	NA	FALSE	
3345t	es	TRUE	
3qq56	nl	FALSE	
44qw3	ol	FALSE	
4550e	NA	FALSE	
3369w	NA	TRUE	
:			
:			

1. **Language filtering:** separate classification for each language
2. **Drop duplicated text id's:** to analyze the same title only once
3. **Correct filter manually:** for projects with statistics = TRUE, some projects manually corrected*

Data set for title text mining

text_id	long_language	statistics	descr2mine
342ex	en	TRUE	The last…
ee455	en	TRUE	Building..
290r4	en	FALSE	A land …
3345t	en	FALSE	Mice…
3369w	en	FALSE	Sexual …
	:		
	:		

learning set

prediction set

Title pattern matching

1. **Construct learning and prediction set:** learning set with 50% TRUE, 50% randomly chosen FALSE projects, rest as prediction set
2. **Clean and lemmatize descr2mine**
3. **Create DTM matrices:**
   3.1 Split up learning set into 80% training data, 20% testing data
   3.2 Create training DTM
   3.3 Create test & prediction DTM with terms that appear in training set
4. **Train the XGBoost model**
5. **Test and predict the XGBoost model:** every project in the prediction set is assigned a probability p_{stat} that it is a statistical project -> if above 0.9, it is classified as statistical
6. **Iterate step 1-5. for learning set robustness:** since 50% as FALSE marked projects are included in the training set, reconstruct learning set with highly improbable stat. projects with $p_{stat} \leq 0.3$

text_id	statistics	p_{stat}	text_mining_stat
322ex	FALSE	0.922	TRUE
ee245	FALSE	0.23	FALSE
290r4	FALSE	0.899	FALSE
3345t	FALSE	0.95	TRUE
3369w	FALSE	0.123	FALSE
	:		
	:		

Add detection filters to original reduced data set acc. to text id

*currently only done for en & 200 projects

4.1 Technical aspects of text mining

PARiS 21

For a **gender classification**, the "statistics" filter on the previous slide is exchanged for the "gender" filter from the title pattern matching. It would also be possible to use a combination of different identifiers such as the gen_donor as the "gender" filter, but a bias can be introduced this way.

The classification is based on a regularizing gradient boosting framework implemented in the **XGBoost** project. For a short but really good introduction into boosted trees, see the project documentation. For more information on gradient boosting, consider this. In this text mining classification, the algorithm learns the importance for each word (e.g. for gender "woman" with very high importance, "air" witch very low importance) and can thereby classify unknown projects based on the words in the description.

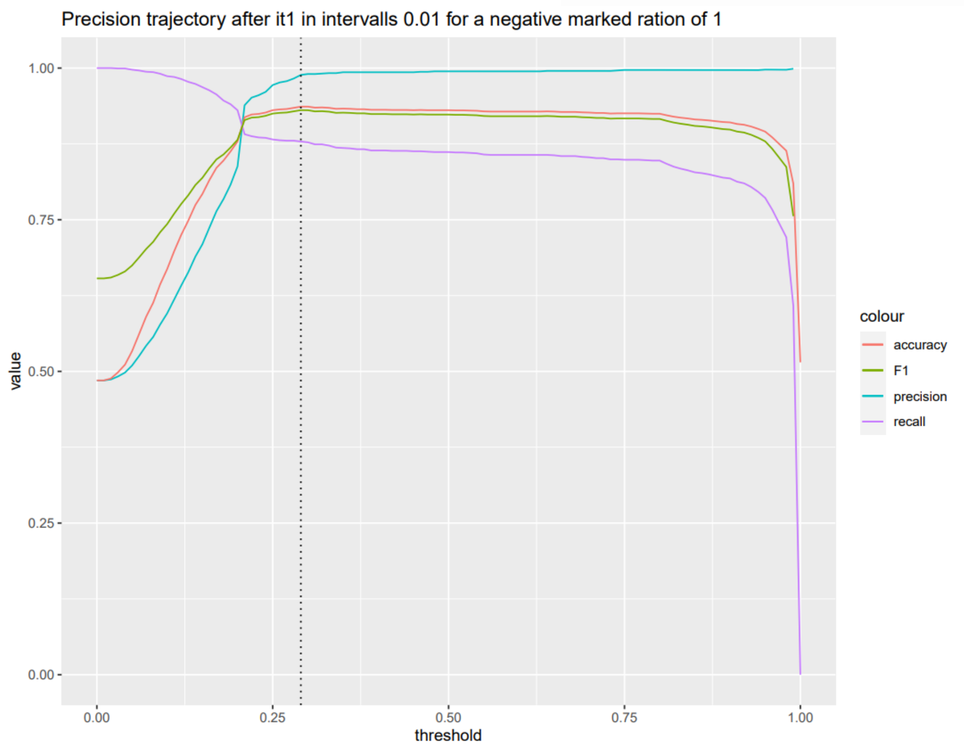
A few **parameters** were introduced for testing purposes or to speed up the process:

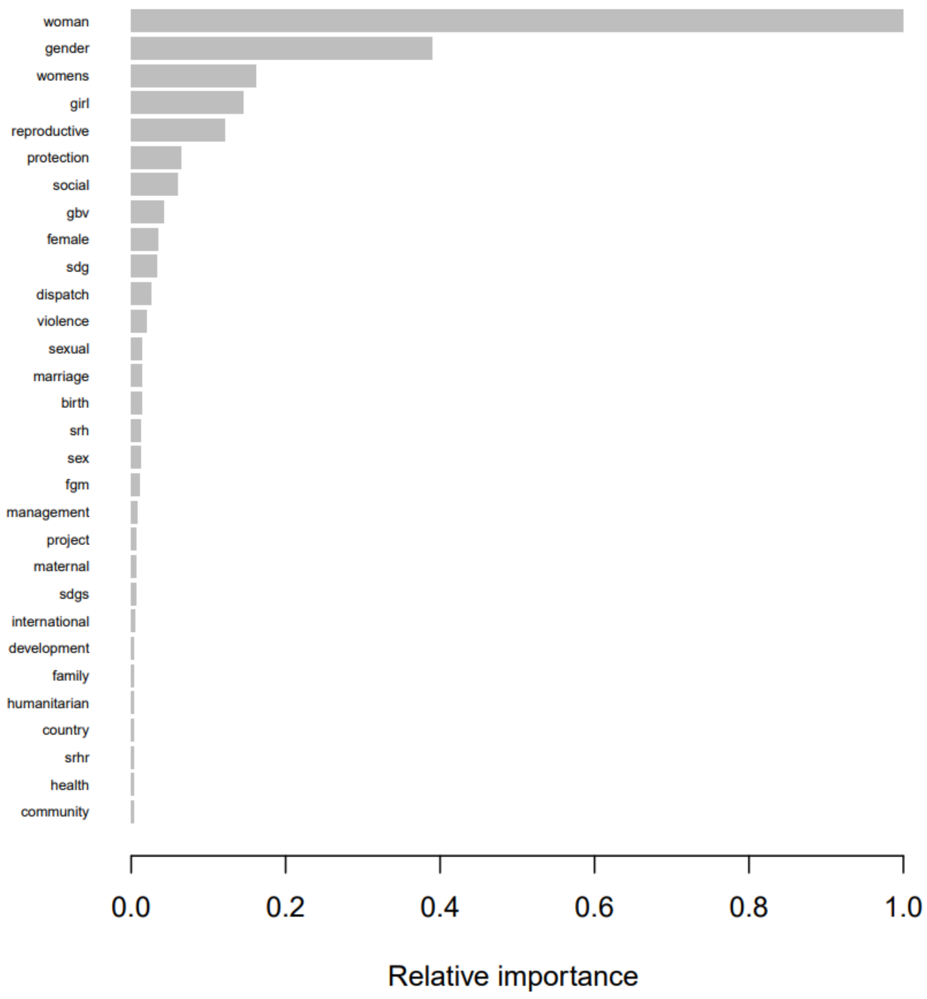
- **n_gram:** if set to integers larger than one, the classification considers also word combinations of that length for the importance matrix (e.g. "information system" with high importance); high computation time for n_gram > 1 due to exponential growth of DTMs
- **full_learning_percent:** take only x% of the full learning set if the size is too large for RAM (only for en & gender)
- **neg_sample_fraction:** Fraction of negatively marked projects to positively marked in the learning set
- **frac_pred_set:** use only x% of full prediction set; can be used to speed up the whole process for testing purposes
- **save_fit_xgb & load_fit_xgb:** can be used to save the fitted model in the second iteration; use load to load a previously fitted model to save computation time
- **split_pred & n_pred_sets:** for splitting up the prediction set into n_pred_sets data frames to handle very large prediction sets; preserves RAM when predictions are made since only a n_pred_setsth of the full prediction set has to be held in the memory at the same time.

4.2 Performance checks

The most insight into the way the classification works is through the **importance matrix.** Through the relative importance of each word, the probability score for a long description can be reconstructed.

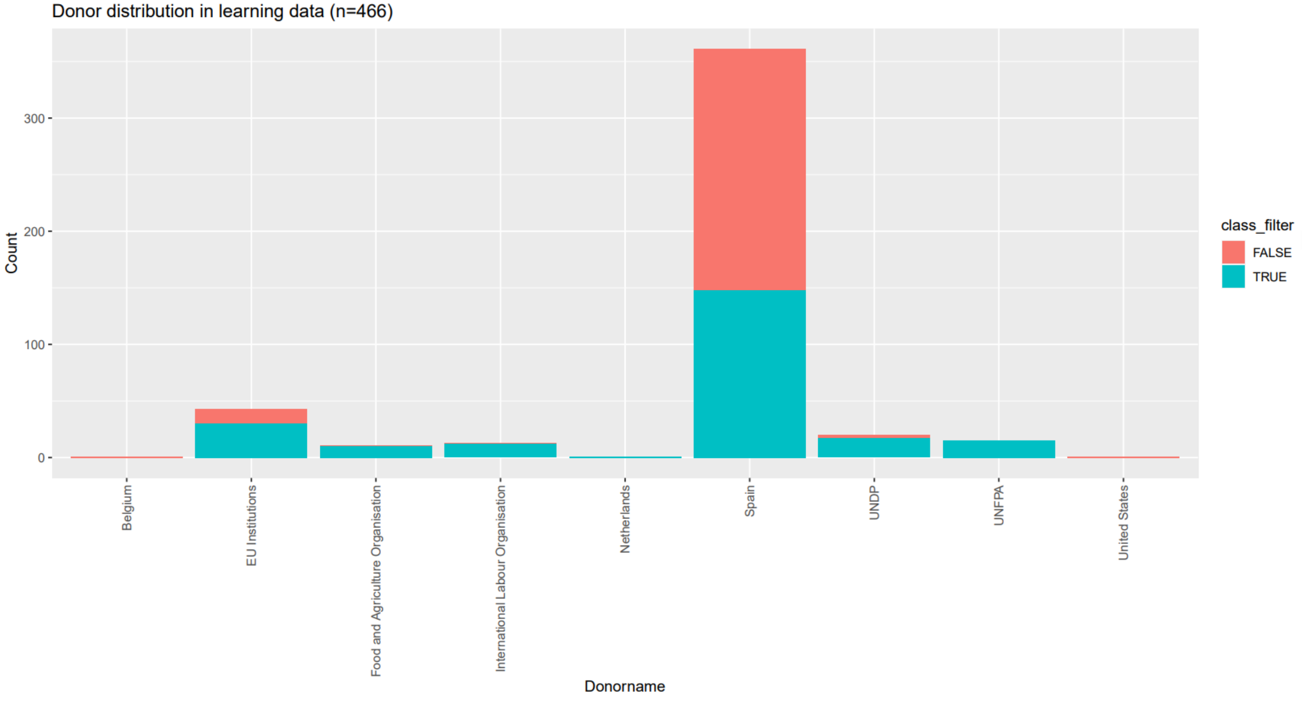
For checking the quality of the classification and determining the optimal threshold, the **precision, accuracy** and **recall** are of interest.





4.2 Performance checks

To avoid biases, it is interesting to check the **donor distribution.** If a certain donor is overrepresented, or in the case of a minority language, a projects in the learning and prediction set stem from a spurious donor, it can be detected.



for Spanish statistical classification