# Multilingual Classification of Economic Activities

23 February 2021

Casper Eriksen, data scientist
ML-Lab

DANISH BUSINESS AUTHORITY

# Agenda

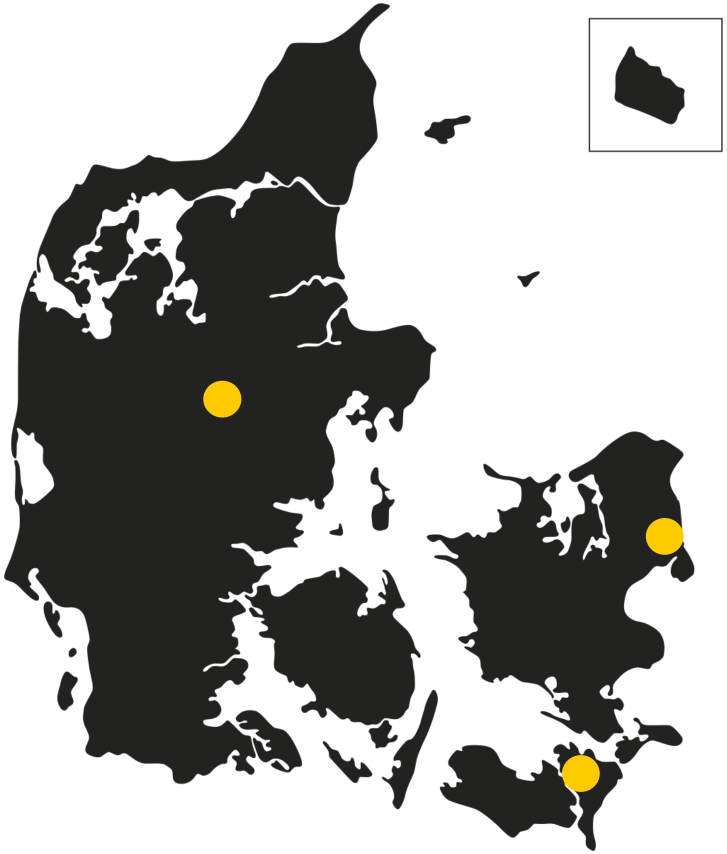
| | |
|---|---|
| 1 | Danish Business Authority |
| 2 | Using ML to Improve the Classification of Danish Businesses |
| 3 | Transfer Learning for Classification of Activities |

# Danish Business Authority

# Danish Business Authority

## …We make it easy and attractive to run a responsible business and create development throughout Denmark

- Good framework conditions for business development
- Effective business regulation and enforcement
- Growth and business development throughout Denmark
- Efficient and professional business service for companies
- Covid-19 compensation schemes



# ML-Lab

- Team of 12 data scientists.

- We build machine learning models.
  - First ML model was put into production in 2017

- We focus on fraud detection with graphs.
  - Graph of 300M nodes and 700M relations

- ...But we also use ML to help businesses.

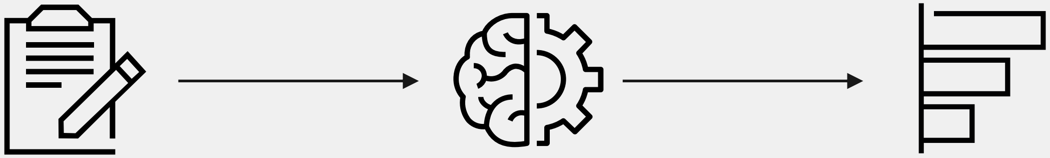
# Using ML to Improve the Classification of Danish Businesses

# Classification of Danish Businesses

- When starting a business in Denmark, the business must select a Danish activity code.
- Denmark uses a subdivision of NACE with 736 national codes (Dansk Branchekode DB07).
- Statistics Denmark has estimated that 20 % of Danish businesses have picked the wrong code.

# Solution: Text Classification

- To help businesses pick the right code, we decided to build a text classifier.

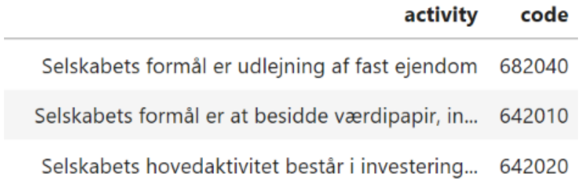


A business writes a short description of its activities

ML model suggests the most likely codes

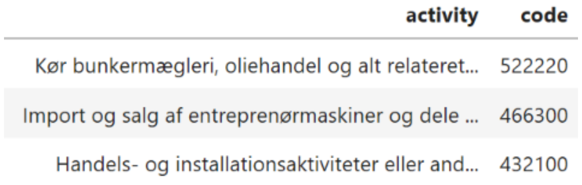
# Data

- 700,000 Danish codes and descriptions of activities extracted from annual reports of Danish companies.



| activity | code |
|---|---|
| Selskabets formål er udlejning af fast ejendom | 682040 |
| Selskabets formål er at besidde værdipapir, in... | 642010 |
| Selskabets hovedaktivitet består i investering... | 642020 |

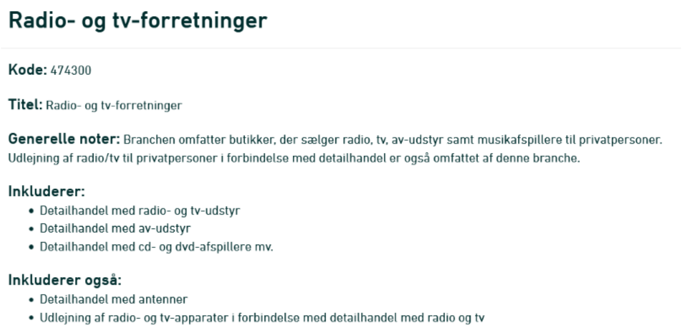
- 1.6M Norwegian codes and descriptions of activities received from Norway’s Brønnøysund Register Centre
  These were machine translated into Danish.



| activity | code |
|---|---|
| Kør bunkermægleri, oliehandel og alt relateret... | 522220 |
| Import og salg af entreprenørmaskiner og dele ... | 466300 |
| Handels- og installationsaktiviteter eller and... | 432100 |

# Synthetic Activities

- For each of the 736 DB07 codes, we generate 200 synthetic activity descriptions by sampling sentences from the official code descriptions.



**Radio- og tv-forretninger**

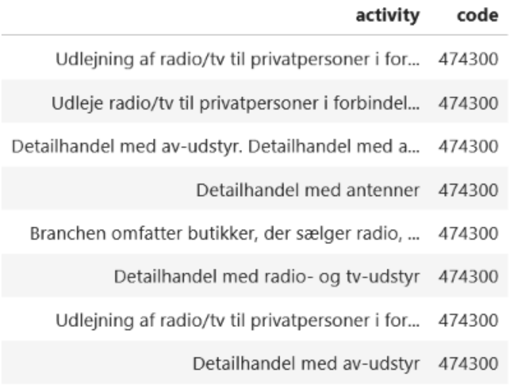
**Kode:** 474300

**Titel:** Radio- og tv-forretninger

**Generelle noter:** Branchen omfatter butikker, der sælger radio, tv, av-udstyr samt musikafspillere til privatpersoner. Udlejning af radio/tv til privatpersoner i forbindelse med detailhandel er også omfattet af denne branche.

**Inkluderer:**
- Detailhandel med radio- og tv-udstyr
- Detailhandel med av-udstyr
- Detailhandel med cd- og dvd-afspillere mv.

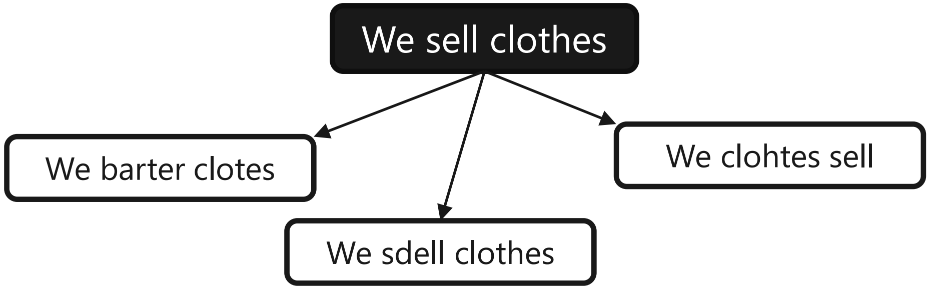
**Inkluderer også:**
- Detailhandel med antenner
- Udlejning af radio- og tv-apparater i forbindelse med detailhandel med radio og tv



| activity | code |
|---|---|
| Udlejning af radio/tv til privatpersoner i for... | 474300 |
| Udleje radio/tv til privatpersoner i forbindel... | 474300 |
| Detailhandel med av-udstyr. Detailhandel med a... | 474300 |
| Detailhandel med antenner | 474300 |
| Branchen omfatter butikker, der sælger radio, ... | 474300 |
| Detailhandel med radio- og tv-udstyr | 474300 |
| Udlejning af radio/tv til privatpersoner i for... | 474300 |
| Detailhandel med av-udstyr | 474300 |

# Text Augmentation

- To increase the robustness of the classifier, all texts are augmented using operations such as:
  - Synonym replacement.
  - Random swap of words or characters.
  - Random deletion of characters.

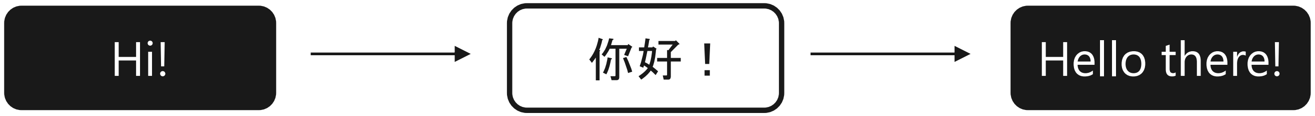


- For this, we've used the Python package **TextAttack**[1].

[1] https://github.com/QData/TextAttack

# Back-Translation

- Back-translation is another text augmentation technique.

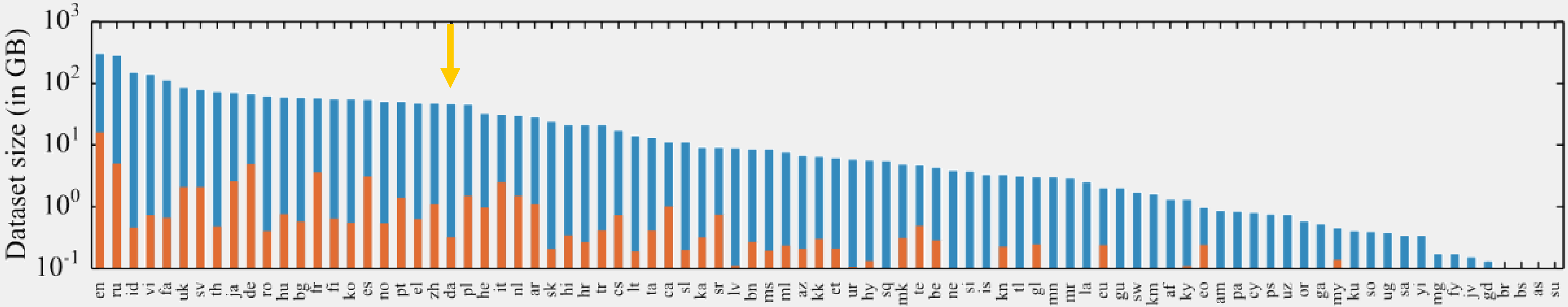


- May be used as a way of generating more data in combination with up-sampling.

# The Model: XLM-RoBERTa

- XLM-RoBERTa[1] is a multilingual transformer-based model pre-trained on 100 different languages.

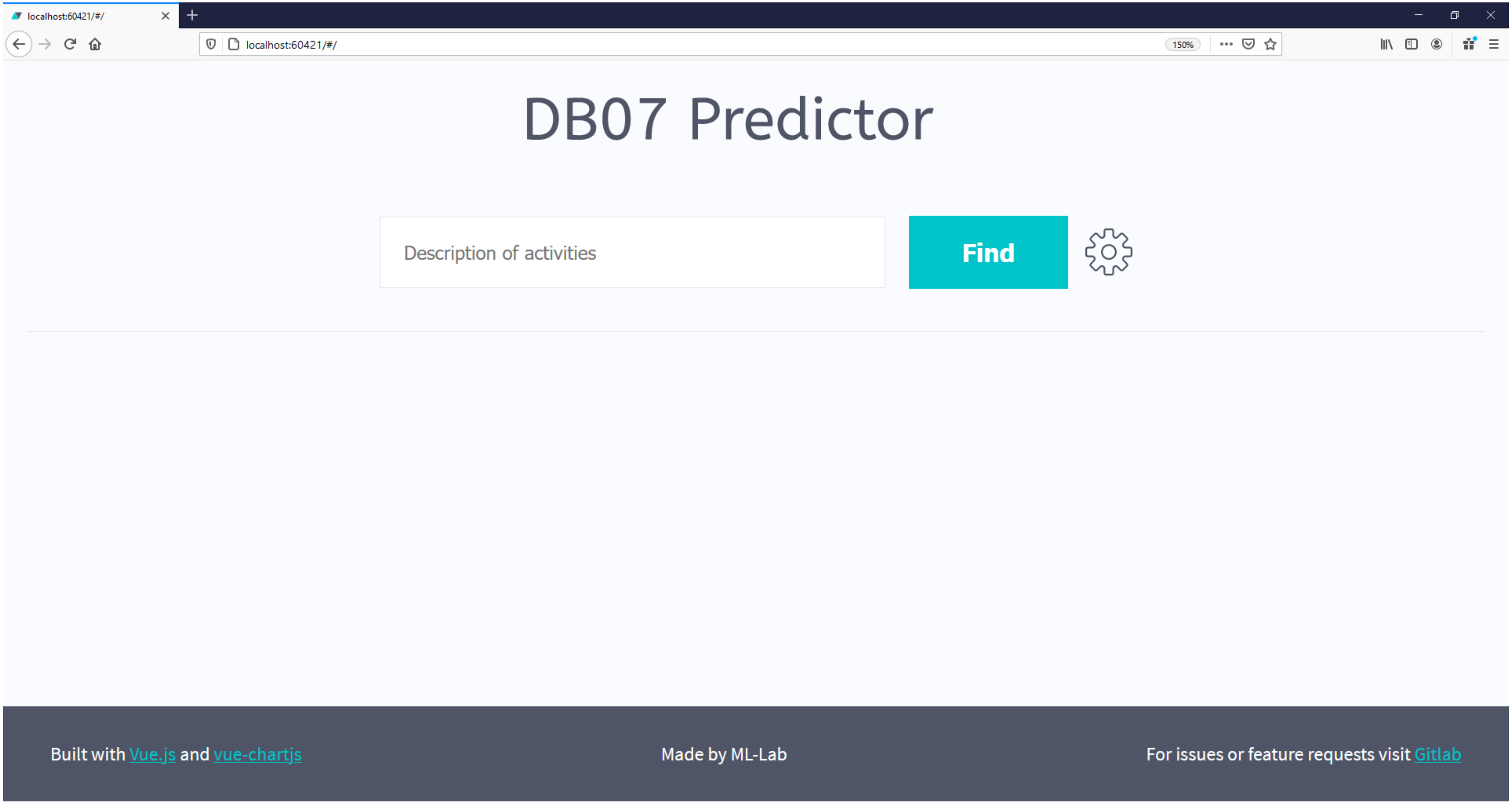
- It has been trained on almost 50 GB of Danish data.
  – more than any other monolingual model.



Amount of data in GB (log-scale) used to train XLM-RoBERTa. See the XLM-RoBERTa paper[2].

[1] https://huggingface.co/transformers/model_doc/xlmroberta.html
[2] https://arxiv.org/abs/1911.02116

# Live Demo



DB07 Predictor

Built with Vue.js and vue-chartjs

Made by ML-Lab

For issues or feature requests visit Gitlab

# Use Cases

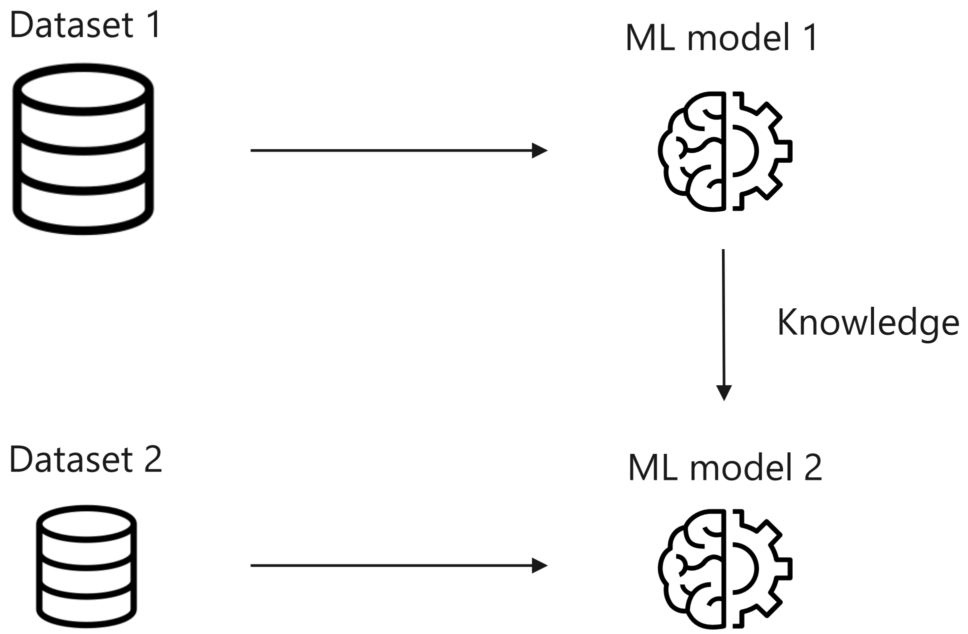
1. Help businesses pick the right code when they register.

2. Identify businesses with wrong activity codes.
   - Compare activity in annual report with selected code.

# Transfer Learning for Classification of Activities

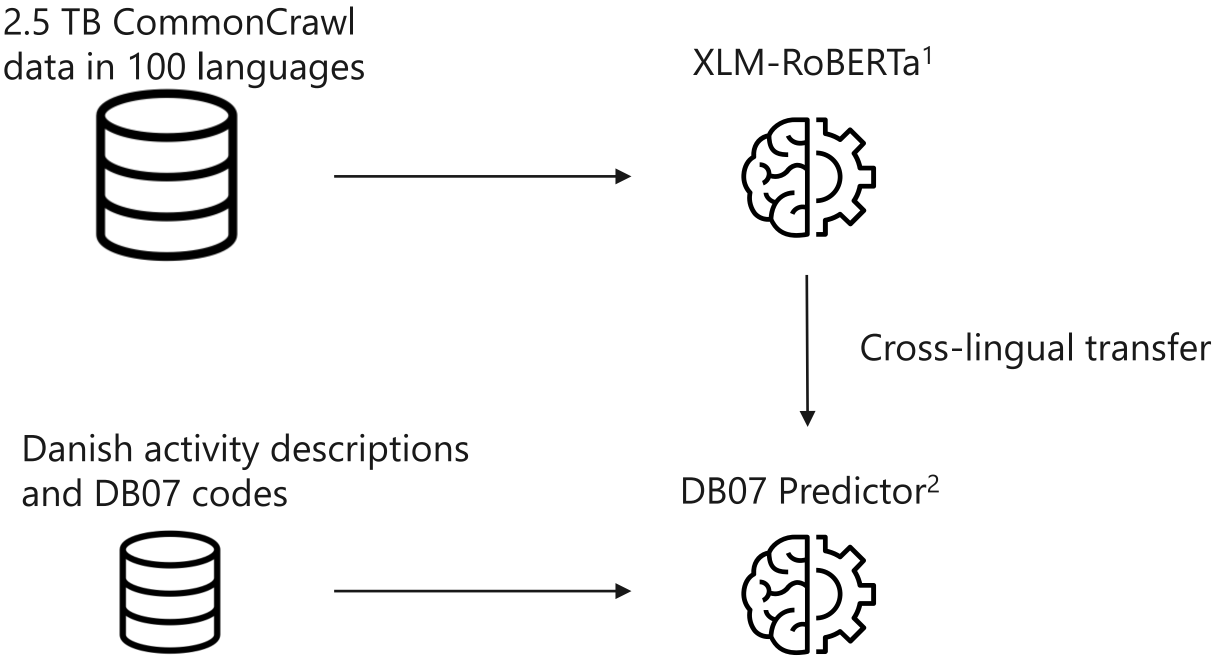
# Transfer Learning

Transferring knowledge from one problem to a different but related problem.



# Transfer Learning

Transferring knowledge of languages to the task of classifying activity descriptions.



2.5 TB CommonCrawl
data in 100 languages

XLM-RoBERTa[1]

Cross-lingual transfer

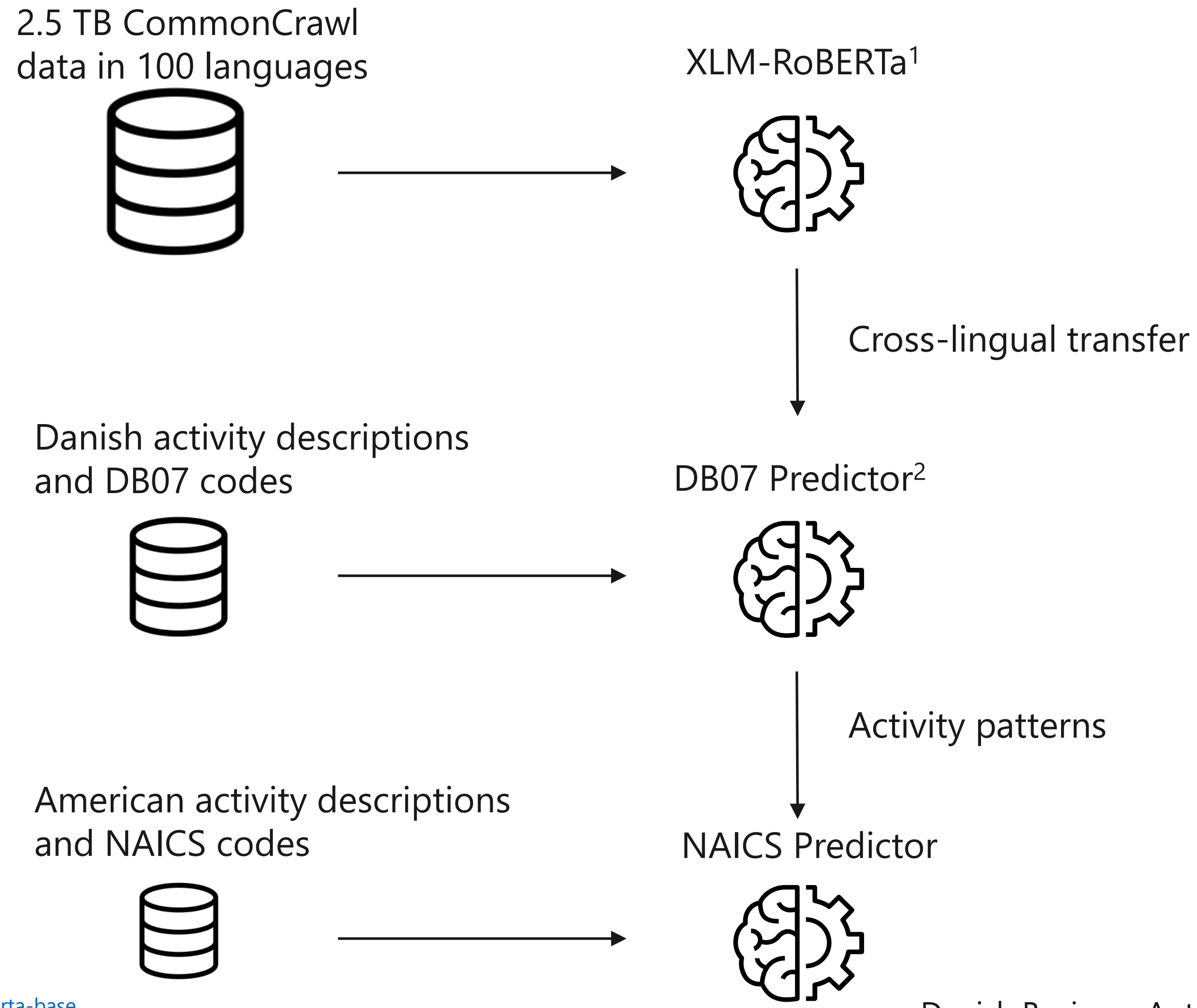
Danish activity descriptions
and DB07 codes

DB07 Predictor[2]

[1] https://huggingface.co/xlm-roberta-base
[2] https://huggingface.co/erst/xlm-roberta-base-finetuned-db07

# Transfer Learning

Transferring knowledge of languages to the task of classifying activity descriptions.



2.5 TB CommonCrawl
data in 100 languages

XLM-RoBERTa[1]

Cross-lingual transfer

Danish activity descriptions
and DB07 codes

DB07 Predictor[2]

Activity patterns

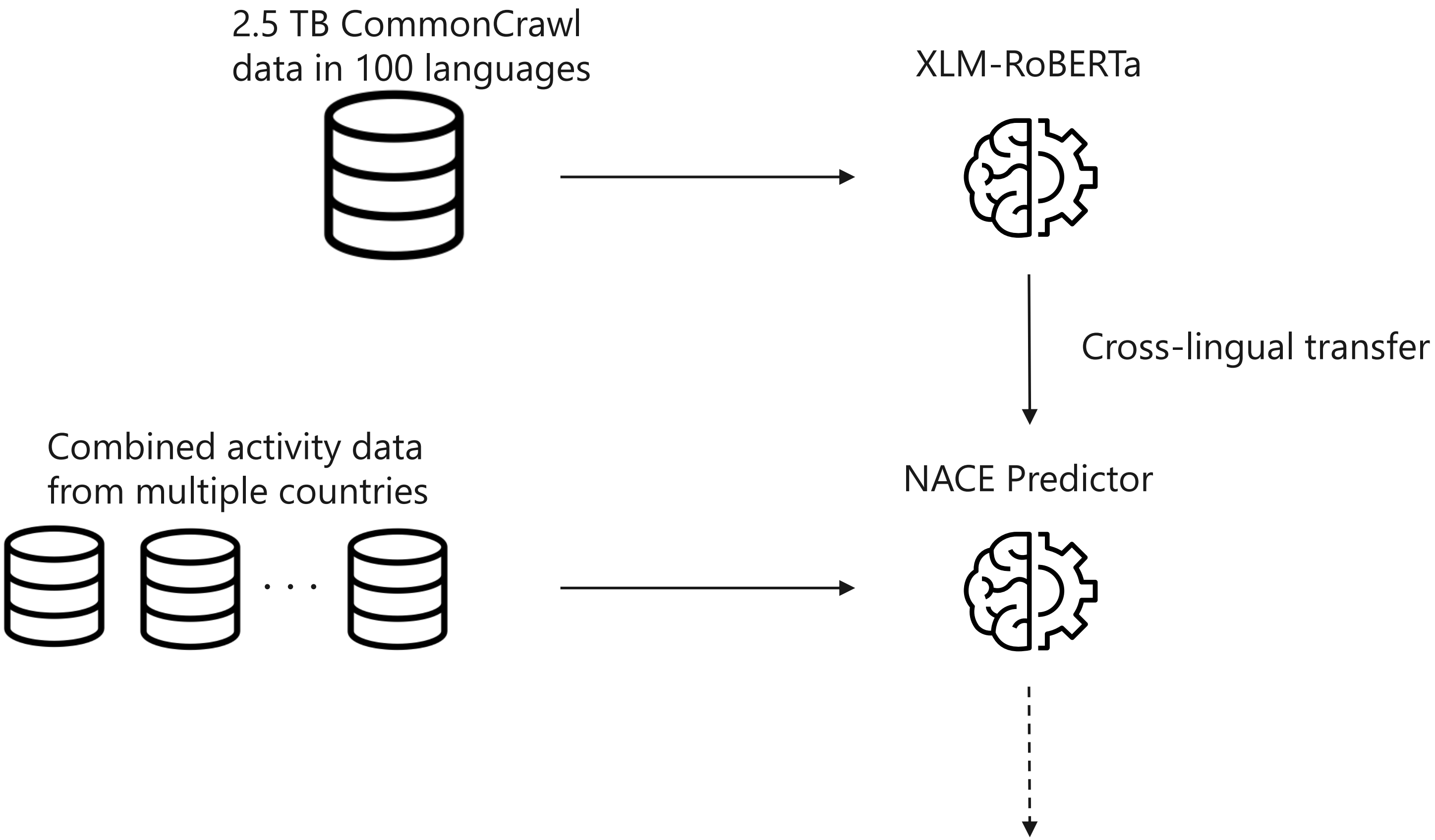
American activity descriptions
and NAICS codes

NAICS Predictor

[1] https://huggingface.co/xlm-roberta-base
[2] https://huggingface.co/erst/xlm-roberta-base-finetuned-db07

# Proposal: Building a Better Basis Model

Collect data from multiple countries to build the best possible basis model for classifying economic activities.



2.5 TB CommonCrawl
data in 100 languages

XLM-RoBERTa

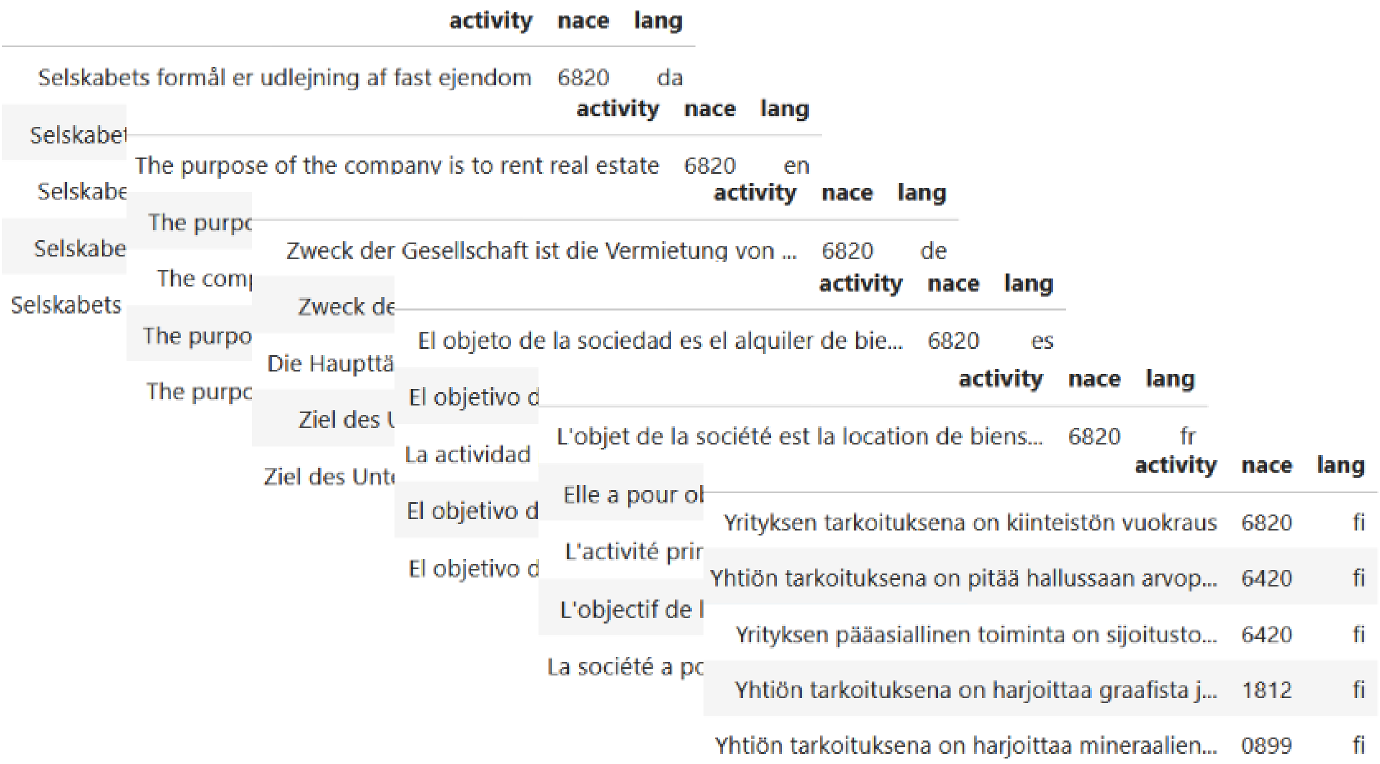
Cross-lingual transfer

Combined activity data
from multiple countries

NACE Predictor

# Improving the Basis with Neural Machine Translation

- Might be unfeasible to obtain data from other countries.
- …Therefore, I'm experimenting with improving the basis by translating the Norwegian and Danish data into other languages using NMT.



| activity | nace | lang |
|---|---|---|
| Selskabets formål er udlejning af fast ejendom | 6820 | da |
| Selskabet | | |
| Selskabe | | |
| Selskabe | | |
| Selskabets | | |

| activity | nace | lang |
|---|---|---|
| The purpose of the company is to rent real estate | 6820 | en |
| The purpc | | |
| The comp | | |
| The purpo | | |
| The purpc | | |

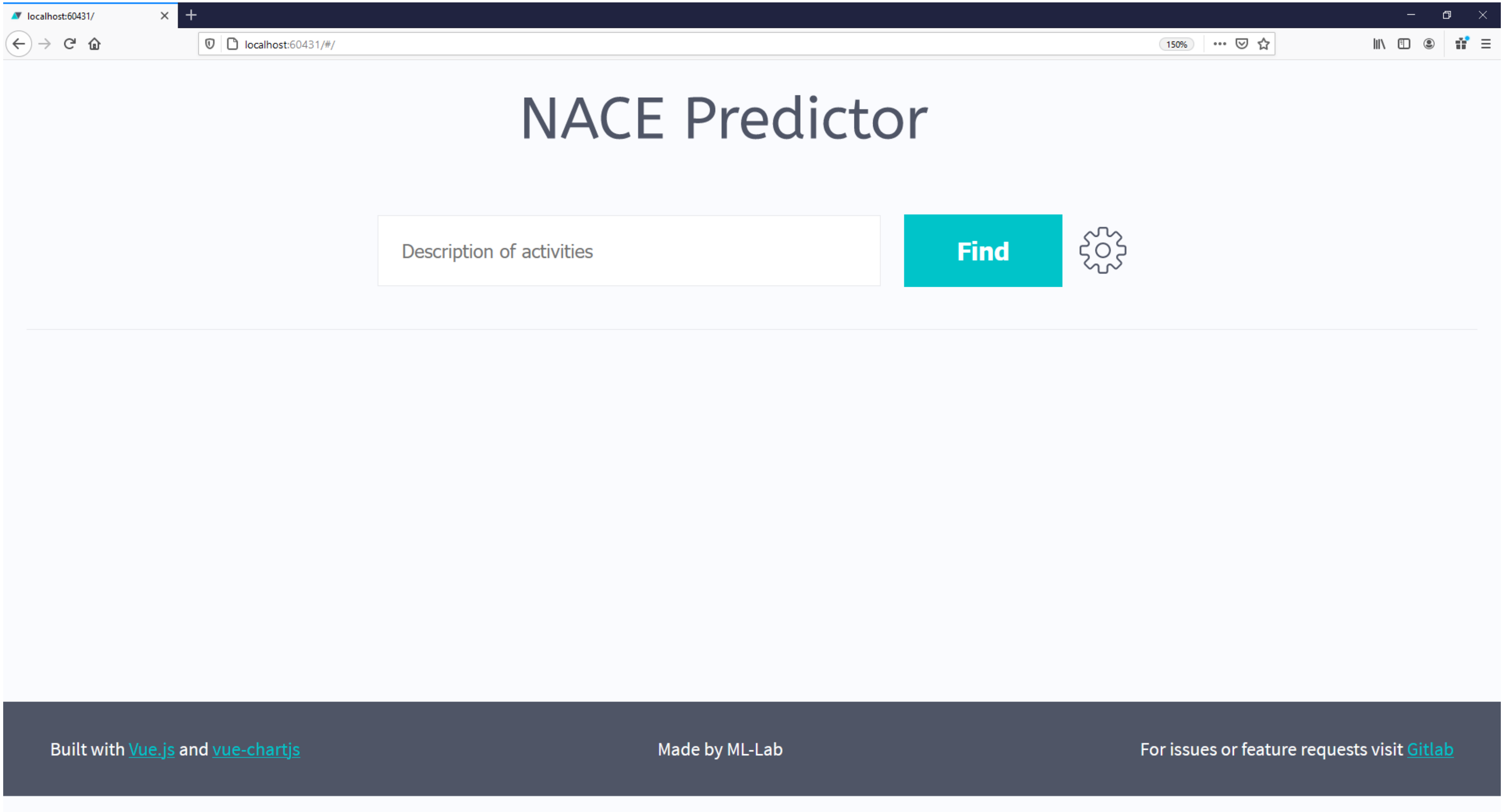
| activity | nace | lang |
|---|---|---|
| Zweck der Gesellschaft ist die Vermietung von ... | 6820 | de |
| Zweck de | | |
| Die Hauptta | | |
| Ziel des U | | |
| Ziel des Unte | | |

| activity | nace | lang |
|---|---|---|
| El objeto de la sociedad es el alquiler de bie... | 6820 | es |
| El objetivo d | | |
| La actividad | | |
| El objetivo d | | |
| El objetivo d | | |

| activity | nace | lang |
|---|---|---|
| L'objet de la société est la location de biens... | 6820 | fr |
| Elle a pour ol | | |
| L'activité prir | | |
| L'objectif de l | | |
| La société a pc | | |

| activity | nace | lang |
|---|---|---|
| Yrityksen tarkoituksena on kiinteistön vuokraus | 6820 | fi |
| Yhtiön tarkoituksena on pitää hallussaan arvop... | 6420 | fi |
| Yrityksen pääasiallinen toiminta on sijoitusto... | 6420 | fi |
| Yhtiön tarkoituksena on harjoittaa graafista j... | 1812 | fi |
| Yhtiön tarkoituksena on harjoittaa mineraalien... | 0899 | fi |

# Live Demo



NACE Predictor

Description of activities

Built with Vue.js and vue-chartjs

Made by ML-Lab

For issues or feature requests visit Gitlab

Casper Eriksen

CasEri@erst.dk

+45 35 29 15 16

Danish Business Authority