Natural Language Processing voor het automatisch classificeren van boorbeschrijvingen

Katrijn Dirix, Hossein Ghorbanfekr,
Pieter Jan Kerstens

23/02/2024

vito

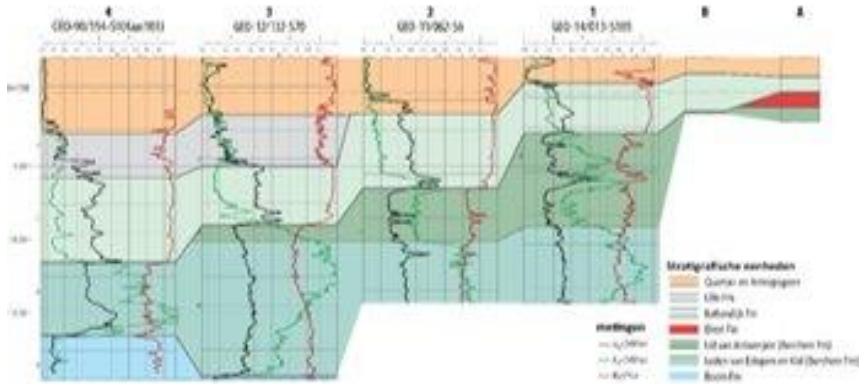
Vlaanderen is milieu

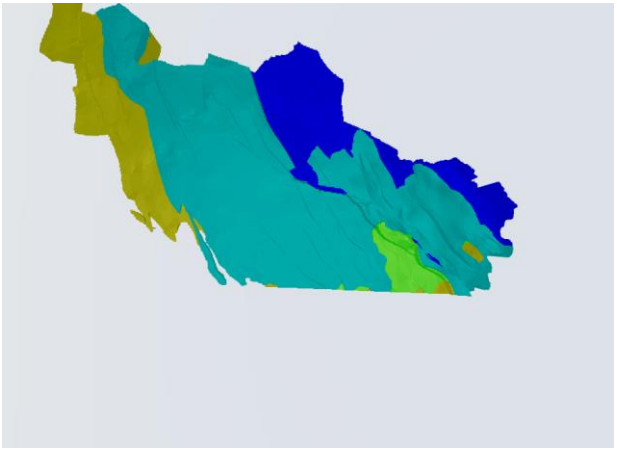
Vlaanderen is omgeving

Context

VLAKO-referentietaak

- **VLAKO:** Vlaams Kenniscentrum Ondergrond
- Uitgevoerd voor het Departement Omgeving (VPO) en de Vlaamse Milieumaatschappij (VMM)
- **Doel**: het ondersteunen van het duurzame beheer van de ondergrond en zijn natuurlijke hulpbronnen
- **Taken**: Lithostratigrafisch onderzoek, beschrijven van ontsluitingen en boringen, opmaken van geologische modellen, ondersteuning van de geotheekwerking, ondersteuning beheer (hydro)geologische data,…

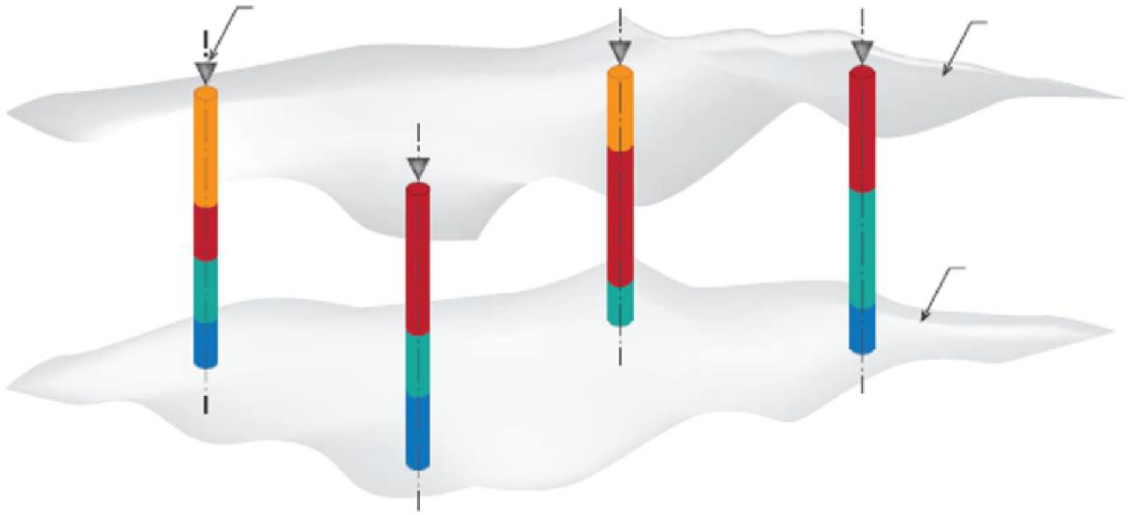


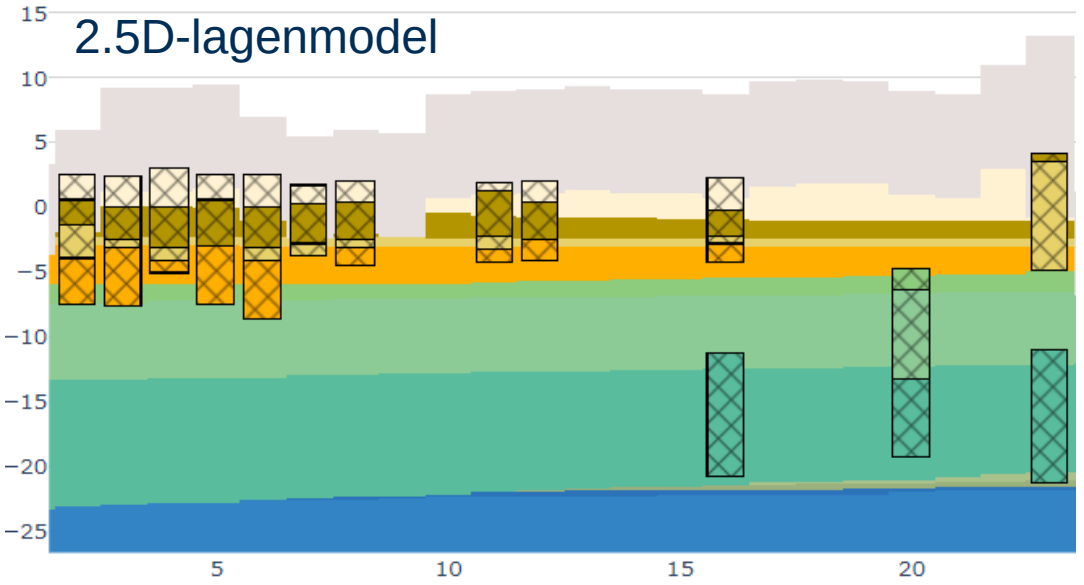


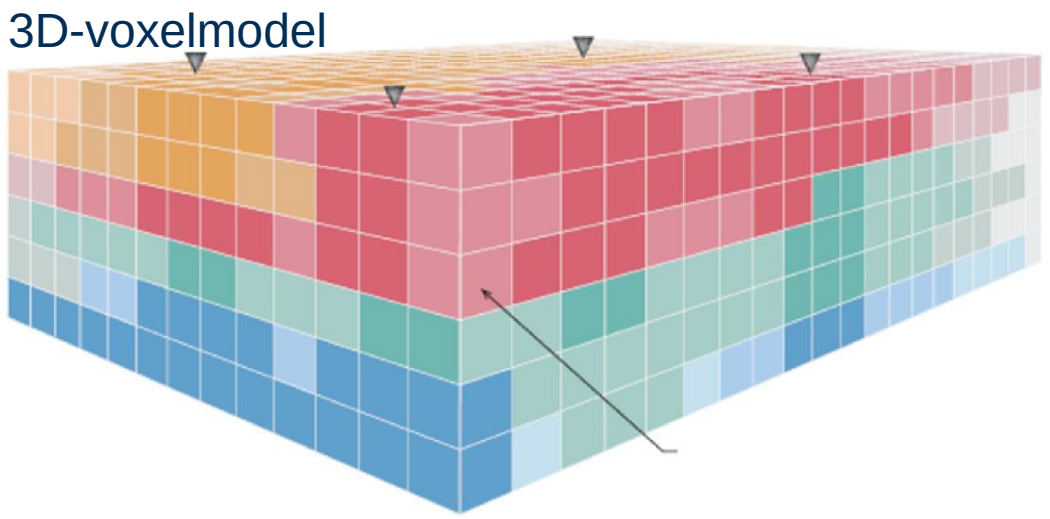
Context

Opmaak 3D geologische modellen.

O.b.v boorbeschrijvingen grondsoort en bijmengingen voorspellen in 3D geologische modellen





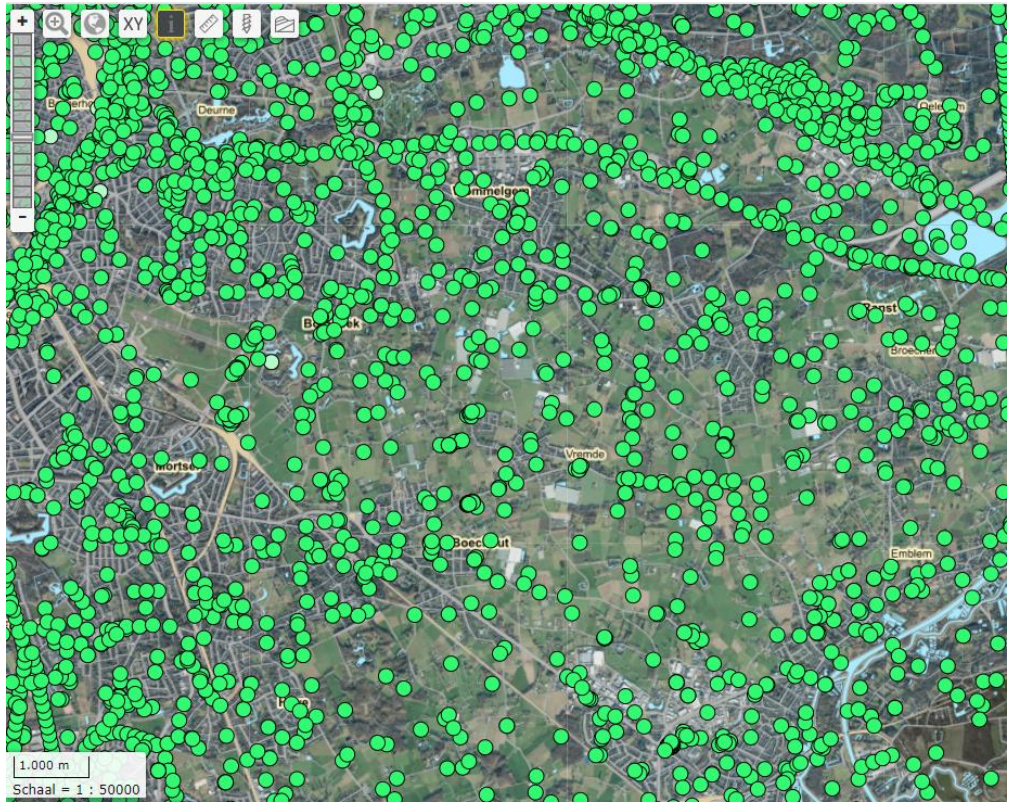


vito

Boorbeschrijvingen als basis

DATABANK ONDERGROND VLAANDEREN

pydov



Lithologische beschrijving - 01/04/1985

Auteur(s): Onbekend (Belgische Geologische Dienst (BGD))

Betrouwbaarheid: onbekend

Kwaliteit:

Score: zeer goed

Boormethode score: Hoog

Van(m)	Tot(m)	M	Beschrijving
0.00	0.39		geroerd
0.39	0.92		grijs leem met oxidatie met veel stippels ijzer/mangaan (0.41 tot 0.56) met laagjes klei (0.56 tot 0.61) graafgang (0.68 tot 0.85)
1.00	1.96		leem met veel homogeen oxidatie (1 tot 1.25) met veel vlekken oxidatie (1.25 tot 1.96) met veel homogeen klei met veel concreties ijzer/mangaan met één niveau zand (1.74 tot 1.79)
2.00	2.57		grijs leem met weinig oxidatie met weinig homogeen klei met concreties kalksteen
2.57	3.60		grijs leem met weinig homogeen klei met lenzen fijn zand met laagjes fijn zand met concreties kalksteen met veel vlekken oxidatie met stippels ijzer/mangaan
3.60	3.74		fijn groen zand met veel homogeen leem met veel glauconiet met één laagjes leem met veel kalksteen
3.74	4.32		grijs leem met weinig homogeen klei met laagjes zand met lenzen zand met concreties kalksteen met één keien silex (3.89 tot 3.89) met oxidatie met veel kalksteen
4.32	4.34		plastisch bruingrijs klei met veel concreties kalksteen
4.34	4.37		fijn zand met oxidatie met veel homogeen leem met veel glauconiet met veel kalksteen
4.37	4.49		bruingrijs leem met weinig homogeen klei met concreties kalksteen met veel laagjes klei met veel kalksteen
4.49	6.00		klei en leem en zand met veel glauconiet met zeer weinig concreties kalksteen met veel kalksteen alternerende gelaagd, golvend tot subhorizontaal gelaagd, verstoord

vito

vito.be

Coderen van boorbeschrijvingen

Individuele manuele verwerking boorbeschrijvingen onmogelijk

Opstellen van rule-based scripting om boorbeschrijvingen in bulk te coderen

Boorbeschrijving	**Gecodeerde lithologie**			
	Hoofdlitho	**Nevenlitho 1**	**Nevenlitho 2**	**rest**
"Bruingrijs fijn zand, matig kleihoudend, met weinig kleine schelpjes"	Fijn zand	Klei	-	schelpen

Coderen van boorbeschrijvingen met rule-based scripting

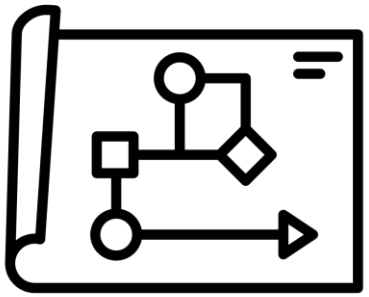
Coderingsmodule

Woordenboek

Woord	type1	type2
zandig	NL	
zand	HL	NL
veen	HL	NL
silt	HL	NL
kleiig	NL	
klei	HL	NL
grind	HL	NL
...	...	...

+

Logische regels



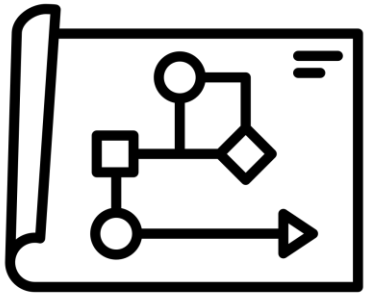
Resultaat

"zand, heel weinig klei, zeer veel keitjes"	zand	klei	grind

Coderen van boorbeschrijvingen met rule-based scripting

Finetune-scripts

Logische regels



+

Regular expressions

```
for woord in categorie:
    lijst = [".?" + j + " [a-z]{0,8}" + woord
             for j in woordenlijst_hoeveelheid]
```

Resultaat

"zand, kleirijk, grofkorrelig, zeer veel keitjes"	~~zand~~	klei	grind

Enter NLP

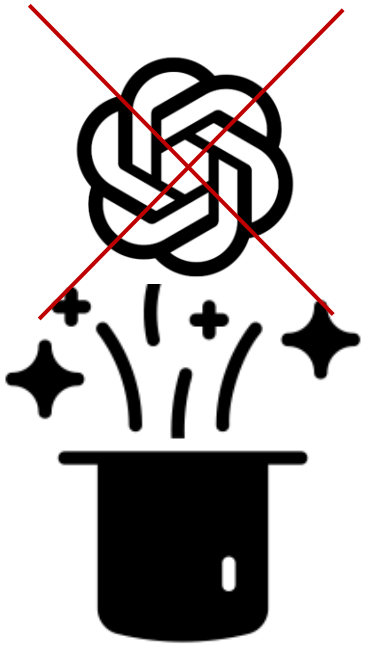
Resultaat: scripts werken prima,
maar bepaalde taalpatronen zijn moeilijk om in logische regels te gieten

Deep-learning Natural Language Processing (NLP) algoritmes als oplossing?

Coderen van boorbeschrijvingen met NLP

Eind 2022: Chat GPT-3: geen goede resultaten:

Hallucinaties…



Boorbeschrijving	Gecodeerde lithologie			
	Hoofdlitho	Nevenlitho 1	Nevenlitho 2	rest
"Bruingrijs fijn zand, matig kleihoudend, met weinig kleine schelpjes"	Fijn zand	Klei	~~grind~~	schelpen

Coderen van boorbeschrijvingen met NLP

Kunnen we zelf een NLP model bouwen, dat specifiek getraind is om (Nederlandse) lithologische beschrijvingen te coderen?

Coderen van boorbeschrijvingen met NLP

Uitdaging: Taalmodellen (bvb. ChatGPT) = Large Language Model (LLM) => Getraind op miljarden teksten => onmogelijk om van scratch op te bouwen

Veel LLM (bvb Google BERT) zijn open source en kunnen gedownload en bijgetraind worden.

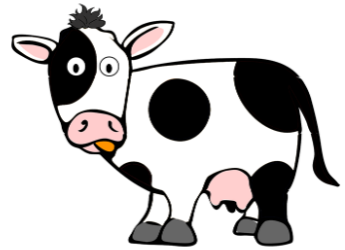
https://github.com/wietsedv/bertje

BERTje: A Dutch BERT model

Wietse de Vries • Andreas van Cranenburgh • Arianna Bisazza • Tommaso Caselli • G Nissim

Model description

BERTje is a Dutch pre-trained BERT model developed at the University of Groninger



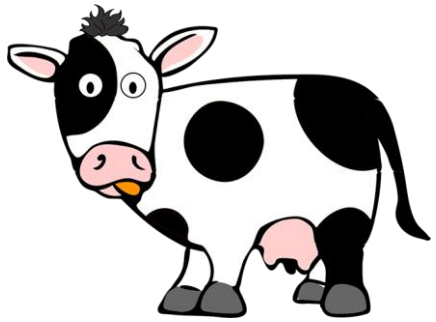
Er bestaat ook een Nederlandstaling LLM, gebouwd volgens de architectuur van BERT: ‘Bertje’

vito

Coderen van boorbeschrijvingen met NLP

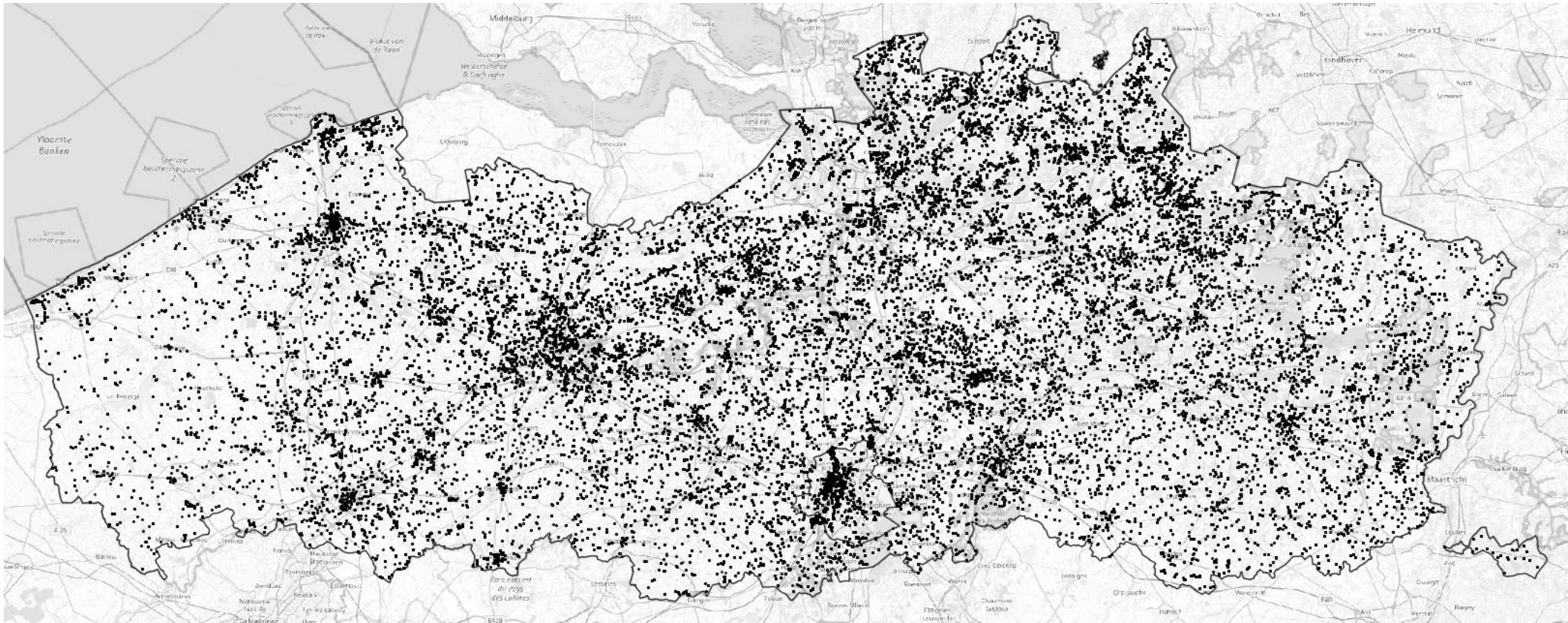
“Transfer-learning” voor NLP-modellen:

Pretraining	Domain adaptation	Fine-tuning
“Language modeling”, begrijpt NL taalpatronen	Raakt vertrouwd met geologische beschrijvingen	Kan beschrijvingen coderen (trainingsdata nodig)



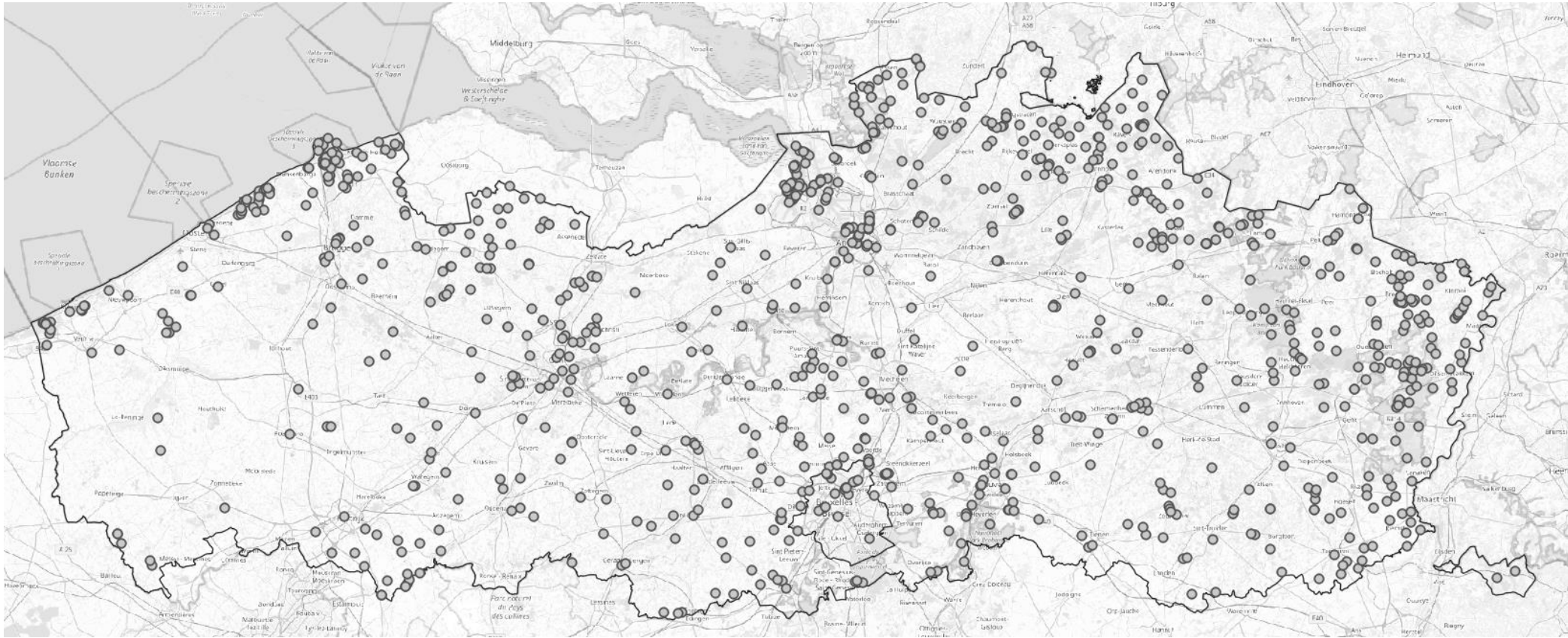
De dataset voor de domain adaptation

- DOV-bevraging: boringen in Vlaanderen met einddiepte > 30m
- +/- 22 000 boorbeschrijvingen
- > 340 000 boorbeschrijving-intervallen



De dataset voor de finetuning

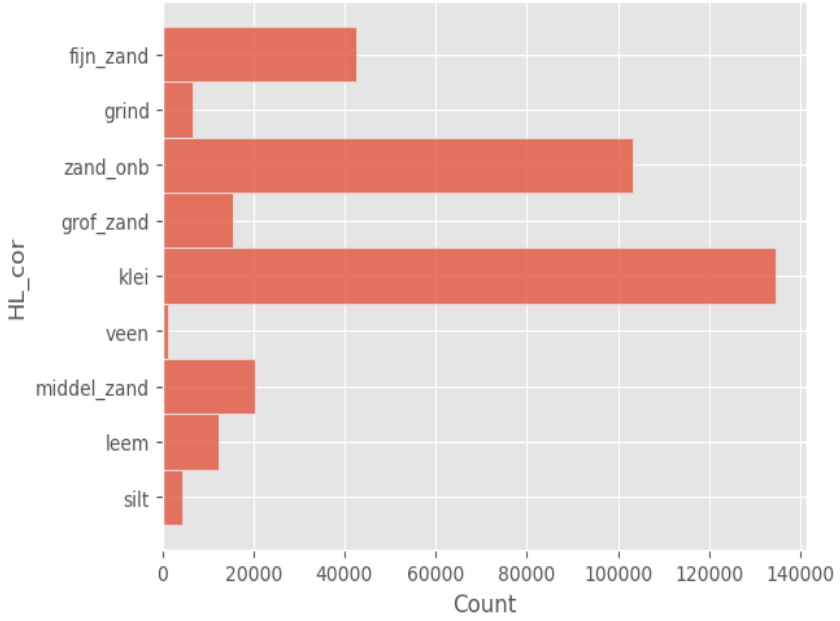
- +/- 2000 manueel gecodeerde boorbeschrijving-intervallen (labels)
- Focus op onderscheid tussen hoofdlithologie (HL) en twee nevenlithologieën (NL1 en NL2)
- Categorieën: zand, fijn zand, middel zand, grof zand, grind, klei, silt, leem, veen

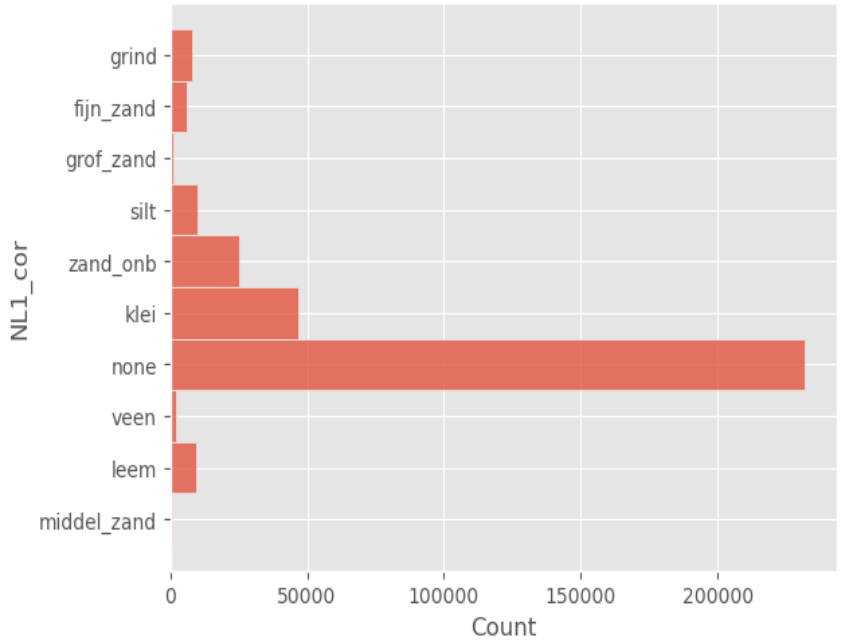


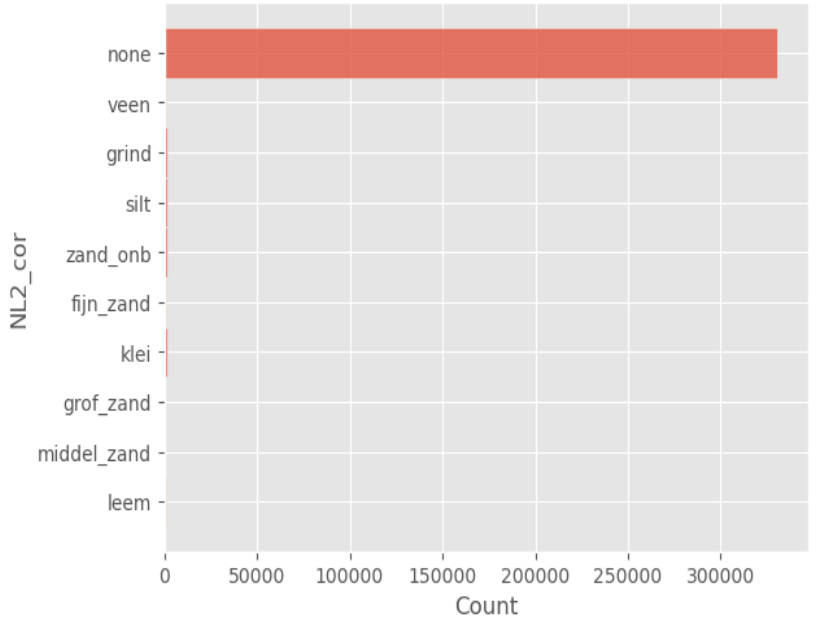
NLP-model

Preprocessing

- ‘idem’ vervangen door voorgaande interval
- Te lange interpretaties elimineren
- Beschrijvingen die meerdere intervallen beschrijven in 1 lijn elimineren
- Dataset is ongebalanceerd => gebruik maken van klassegewichten



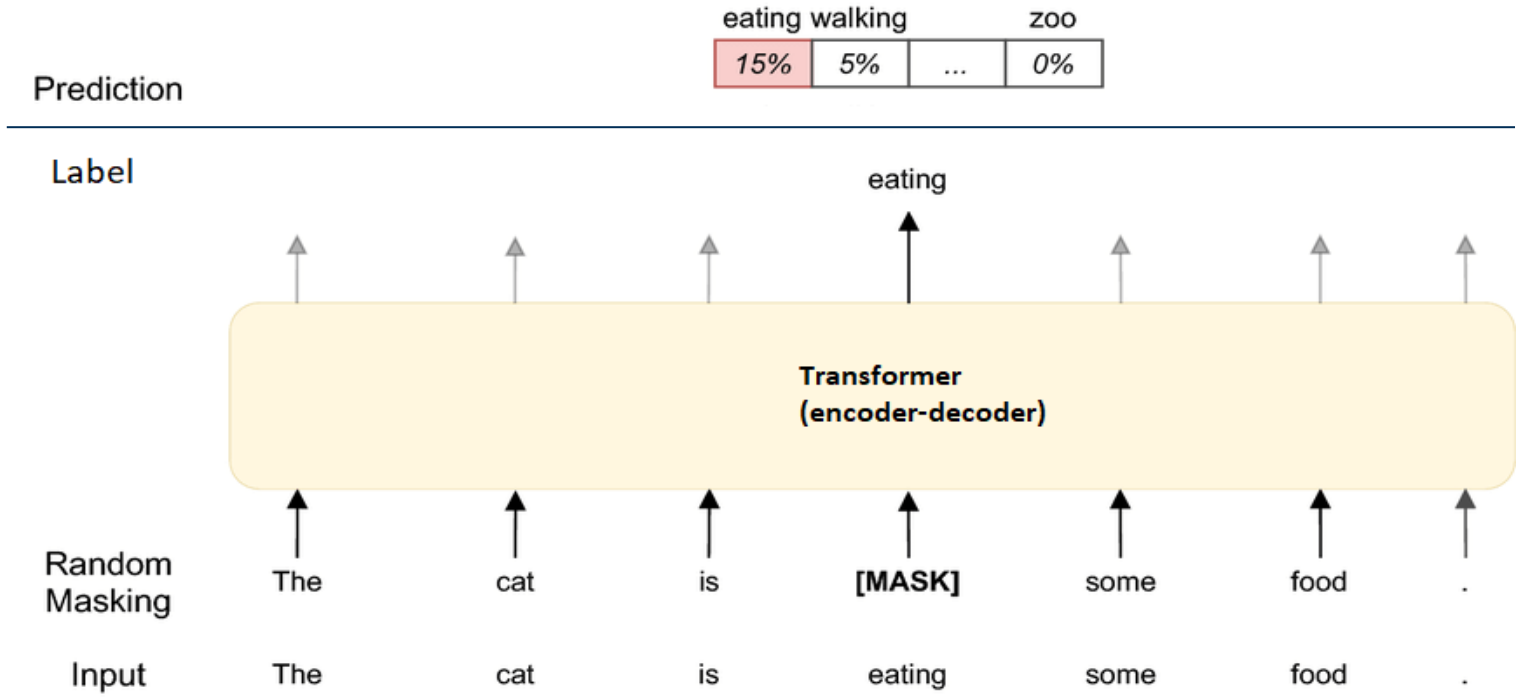




NLP-model

Domain adaptation

- Input = 340k niet geclassificeerde boorbeschrijving-intervallen
- Pre-training = specifieke woorden worden gemasked, model moet het verborgen woord voorspellen

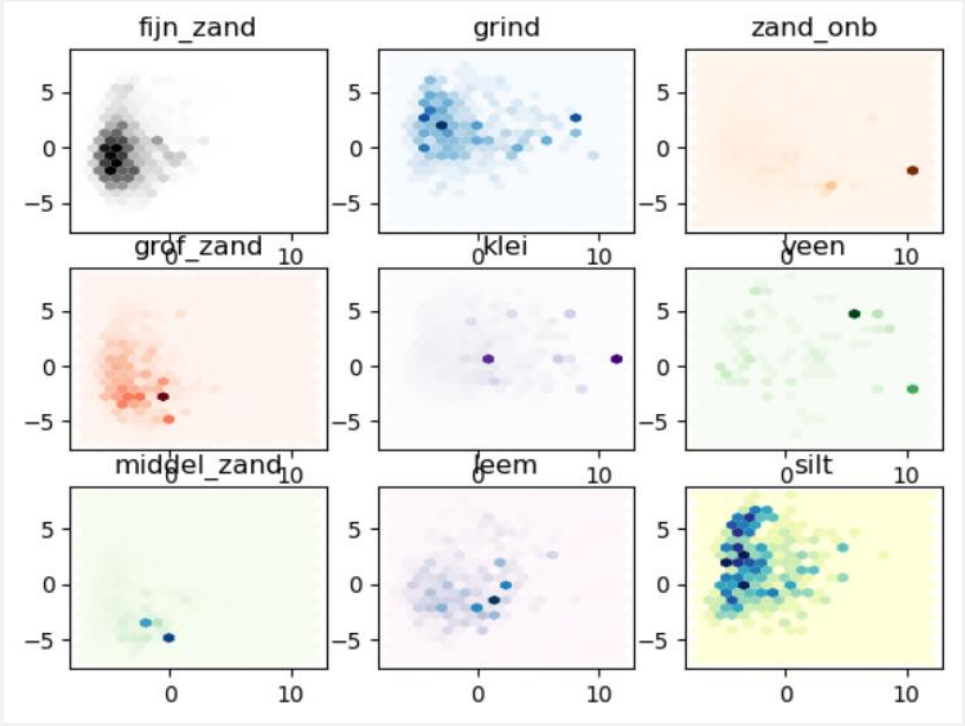


NLP-model

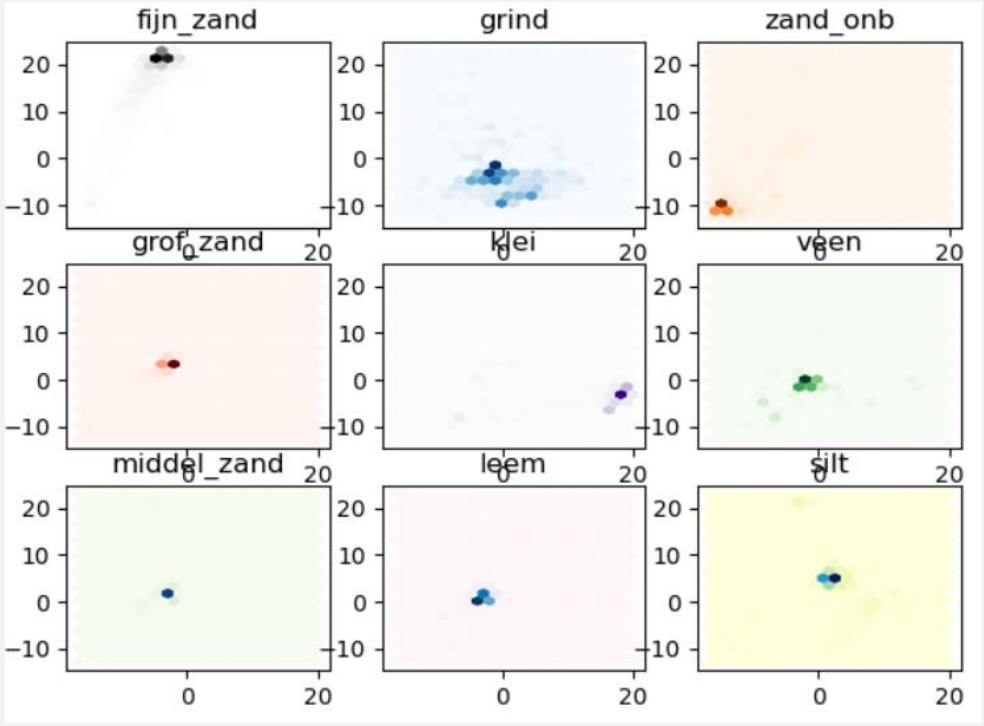
Fine tuning

- Input = 2k manueel geclassificeerde boorbeschrijving-intervallen
- Training = model leert om juiste hoofd- en nevenlithologieën te voorspellen door gebruik te maken van de gelabelde dataset

Pretrained model



Fine-tuned model



NLP-model

Resultaat

FastAPI 0.1.0 OAS 3.1

/openapi.json

"Bruingrijs fijn zand, matig kleihoudend, met weinig kleine schelpjes"

default

GET /predict Drill Core Classes

Parameters Cancel

Name	Description
dutch_description * required string (query)	Bruingrijs fijn zand, matig kleihoudend, met w

Code Details

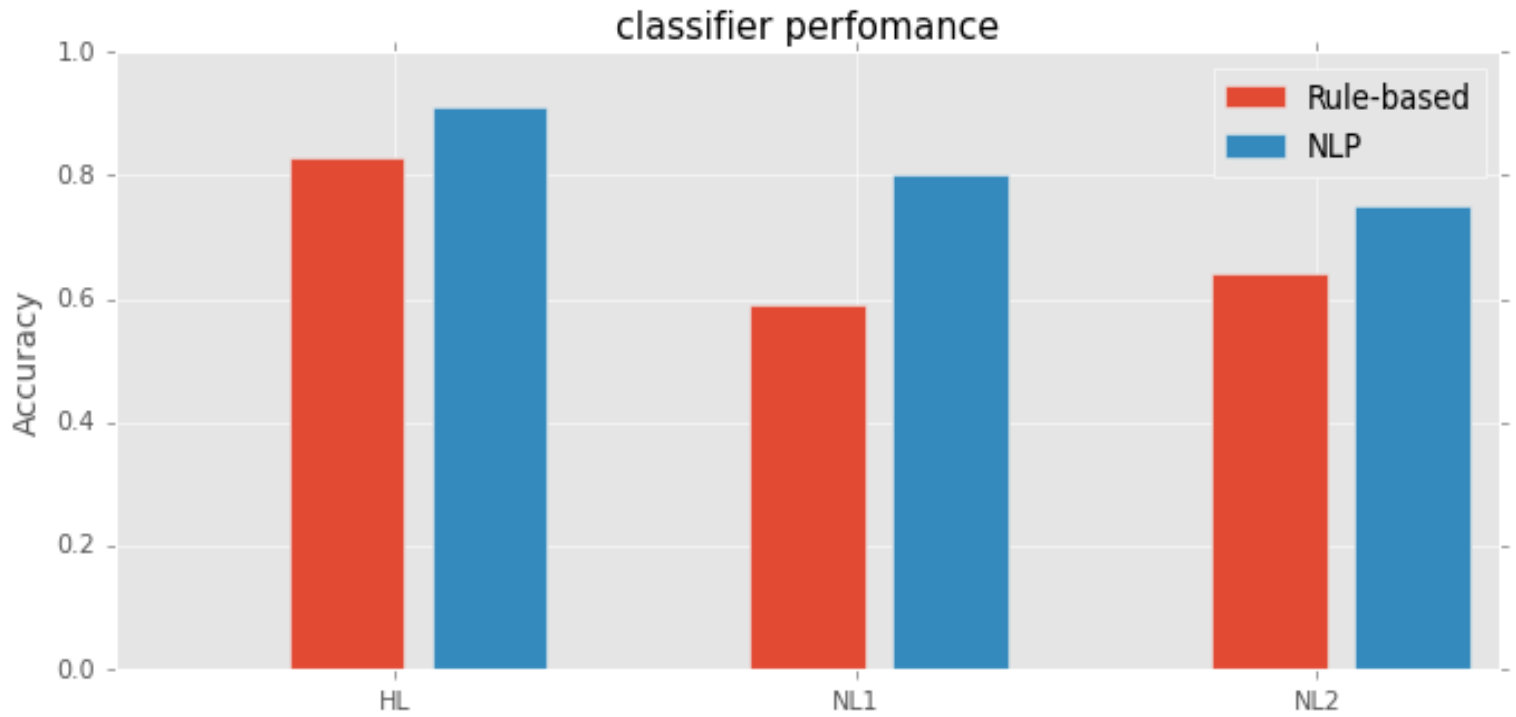
200

Response body


```
{
  "HL_cor": [
    {
      "label": "fijn_zand",
      "score": 0.9935482144355774
    }
  ],
  "NL1_cor": [
    {
      "label": "klei",
      "score": 0.9550267457962036
    }
  ],
  "NL2_cor": [
    {
      "label": "none",
      "score": 0.6851084232330322
```

NLP-model

Resultaat



- Codering o.b.v. NLP gemiddeld beter dan rule-based
- Individuele verschillen tussen klasses
- Work in progress!

NLP-model

Vervolgstappen

- Testen en analyse van NLP-resultaten

- Momenteel aparte modellen voor HL, NL1 en NL2 => modellen koppelen

En Chat GPT-4?

- GPT-4 is veel beter in codering dan GPT-3
- Niet open source => via prompt engineering + API-key classificatie uitvoeren

Pro GPT-4	Con GPT-4
Zeer toegankelijk (!)	Niet open source (!)
Degelijke resultaten	Niet reproduceerbaar
	Hallucinaties (vooral bij klasse zand)
	Zeer traag, en data moet worden opgesplitst
	Postprocessing nodig

Bedankt!

katrijn.dirix@vito.be