# BERT's little minion:

## A tool to normalise time periods extracted from Dutch text

A. Brandsen

Oxford / Online, 10-08-2022

Universiteit Leiden
The Netherlands

AGNES

# Hello!

I am Alex Brandsen

Postdoc at the Faculty of Archaeology

🐦 @alex_brandsen

# BERT's little minion:

## A tool to normalise time periods extracted from Dutch text

A. Brandsen

Oxford / Online, 10-08-2022

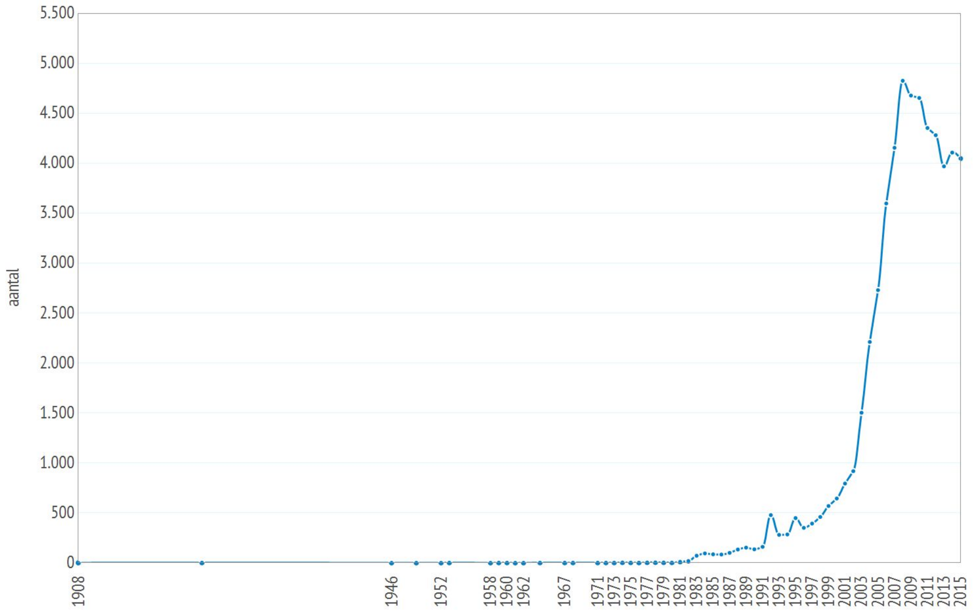
Universiteit Leiden
The Netherlands

AGNES

# Problem: too much text

- Development-led commercial archaeology due to Malta treaty
- More than 80k reports available, growing with ~8k a year
- Search possible on metadata

Archeologische onderzoeksmeldingen in Nederland - 1908-2015



Eenheid: aantal

Bron: RCE - Monumentenregister - ARCHIS

# FULL TEXT SEARCH

Works, but doesn’t account for synonymy and polysemy

## Synonymy: Neolithic

- Neolithic
- Late Stone Age
- 3500 BC
- 5000 - 4000 BP
- 4th Millenium BC
- And so on…

## Polysemy: Swifterbant

- Place
- Excavation event
- Pottery type
- Time period

# Named Entity Recognition (NER)

… the excavation in Amsterdam was …

… fragments of pottery nearby …

… dated to the Iron Age. Other finds …

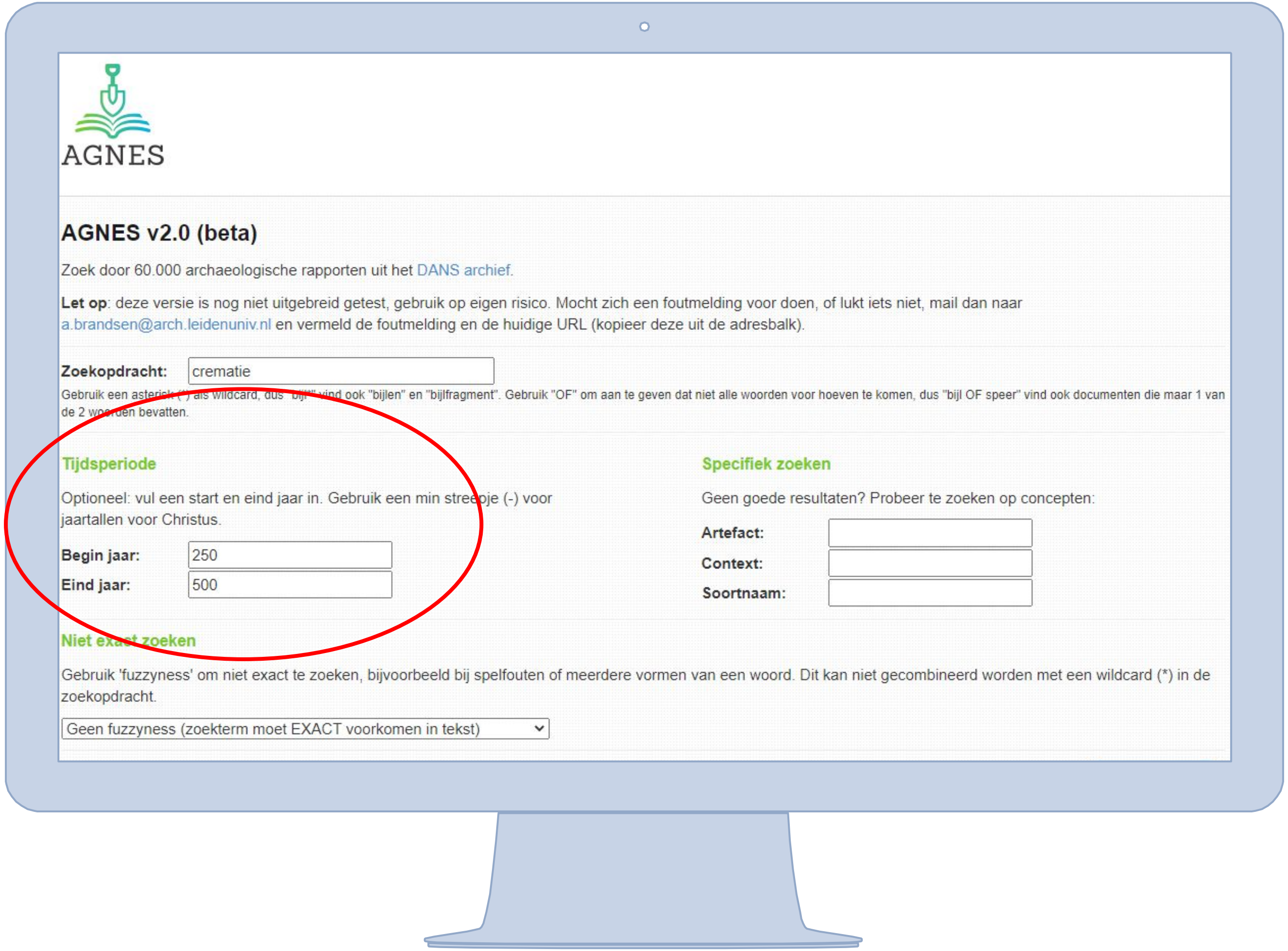
# Deep Learning with BERT

(Bidirectional Encoder Representations from Transformers)

- BERT is a SotA deep learning technique (by Google)
- It can 'learn' a language by looking at large amounts of text
- Afterwards, you can teach it to classify, using human-labelled data
- Created ArcheoBERTje, taught it to detect entities
- Accuracy: 74% (Strict micro F1 score over all classes)

# Labelling Full Document Collection

| Entity | Total | Unique | Top 5 |
|---|---:|---:|---|
| Artefacts | 2,520,492 | 53,675 | pottery, charcoal, flint, bone, brick |
| Contexts | 1,602,124 | 21,319 | pit, ditch, posthole, well, house |
| Materials | 457,031 | 6,146 | wooden, flint, wood, metal, bronze |
| Locations | 3,488,698 | 147,077 | nederland, ' , groningen, noord - brabant, gelderland |
| Species | 928,437 | 34,540 | cow, hazel, sheep, goat, pig |
| Time Periods | 4,698,323 | 98,445 | roman period, iron age, 150 - 210, late medieval, modern |
| Total | 13,695,105 | 361,202 | |



AGNES

## AGNES v2.0 (beta)

Zoek door 60.000 archaeologische rapporten uit het DANS archief.

**Let op**: deze versie is nog niet uitgebreid getest, gebruik op eigen risico. Mocht zich een foutmelding voor doen, of lukt iets niet, mail dan naar a.brandsen@arch.leidenuniv.nl en vermeld de foutmelding en de huidige URL (kopieer deze uit de adresbalk).

**Zoekopdracht:** crematie

Gebruik een asterisk (*) als wildcard, dus "bijl*" vind ook "bijlen" en "bijlfragment". Gebruik "OF" om aan te geven dat niet alle woorden voor hoeven te komen, dus "bijl OF speer" vind ook documenten die maar 1 van de 2 woorden bevatten.

### Tijdsperiode

Optioneel: vul een start en eind jaar in. Gebruik een min streepje (-) voor jaartallen voor Christus.

**Begin jaar:** 250

**Eind jaar:** 500

### Specifiek zoeken

Geen goede resultaten? Probeer te zoeken op concepten:

**Artefact:**

**Context:**

**Soortnaam:**

### Niet exact zoeken

Gebruik 'fuzzyness' om niet exact te zoeken, bijvoorbeeld bij spelfouten of meerdere vormen van een woord. Dit kan niet gecombineerd worden met een wildcard (*) in de zoekopdracht.

Geen fuzzyness (zoekterm moet EXACT voorkomen in tekst)

# The Minion: extracting year ranges

https://github.com/alexbrandsen/timeperiod2daterange

- Rules based script: convert entity to start year / end year
- ‘Neolithic’ → -5200 / -3000 (using thesaurus)
- ‘1250 BC’ → -1250 / -1250
- ‘second half 3rd century’ → 250 / 300
- ‘1450 ± 50 BP’ → 450 / 550
- ‘Third quarter of the 2nd century to the late Medieval period’ → 150 / 1450

# Extracting Year Ranges

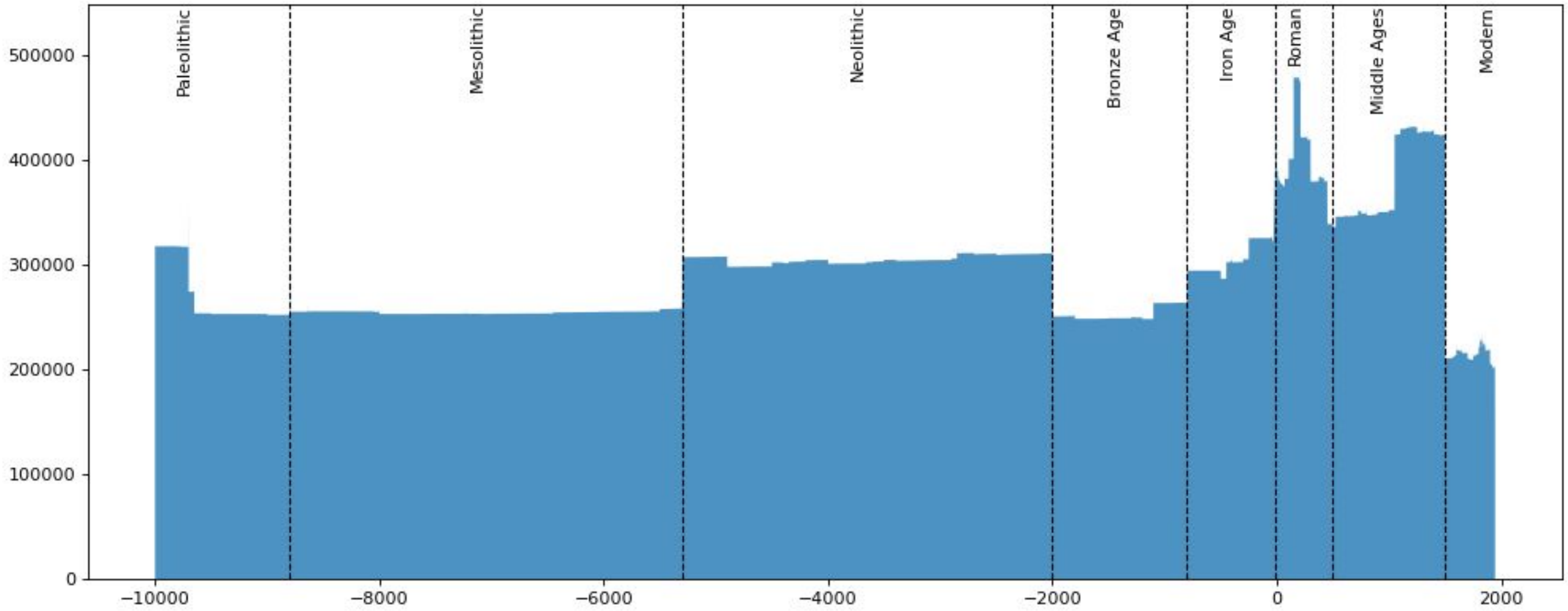
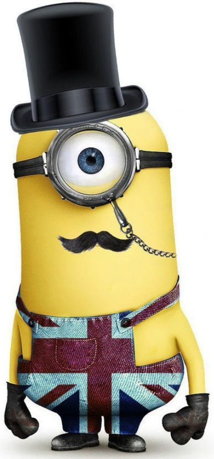


Figure from: Brandsen, A., Verberne, S., Lambers, K., & Wansleeben, M. (2021). Can BERT Dig It? - Named Entity Recognition for Information Retrieval in the Archaeology Domain. *Journal on Computing and Cultural Heritage*. https://doi.org/10.1145/3497842

# Future Work

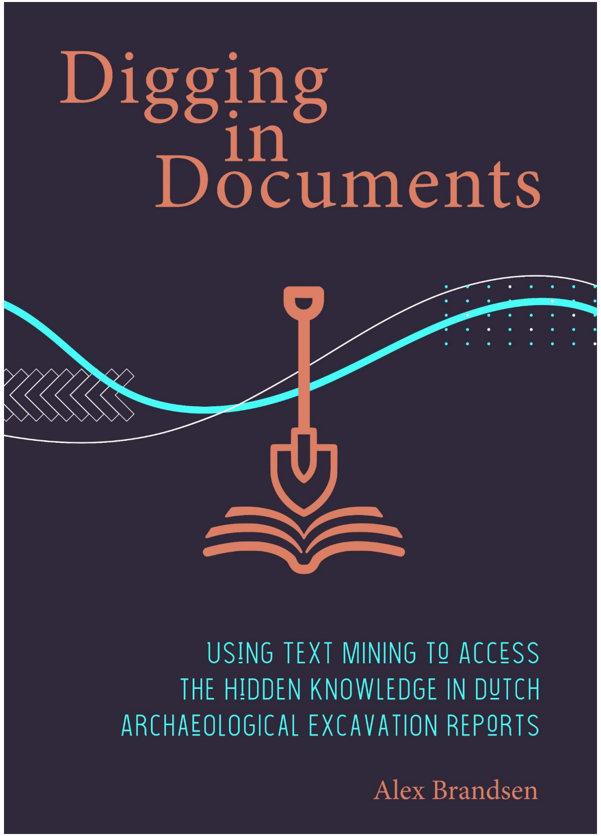
- Multilingual:
  - German
  - English
  - (maybe French?)
- Allows for semantic search:
  - ‘Neolithic’ also finds ‘*Neolithicum*’ and ‘*Kugelamphoren-Kultur*’



# Thanks!

**Any questions?**

a.brandsen@arch.leidenuniv.nl
alexbrandsen.nl



Digging in Documents

USING TEXT MINING TO ACCESS THE HIDDEN KNOWLEDGE IN DUTCH ARCHAEOLOGICAL EXCAVATION REPORTS

Alex Brandsen

https://hdl.handle.net/1887/3274287