

A comparison of machine learning and rule-based approaches for text mining in the archaeology domain, across three languages

A. Brandsen, A. Vlachidis & A. Lien-Talks

CAA 2023 - S22



**Universiteit
Leiden**
The Netherlands





Hello!

I am Alex Brandsen
Postdoc at the Faculty of Archaeology

 @alex_brandsen

A comparison of machine learning and rule-based approaches for text mining in the archaeology domain, across three languages

A. Brandsen, A. Vlachidis & A. Lien-Talks

CAA 2023 - S22



**Universiteit
Leiden**
The Netherlands



Joint work with:



Dr. Andreas Vlachidis
UCL

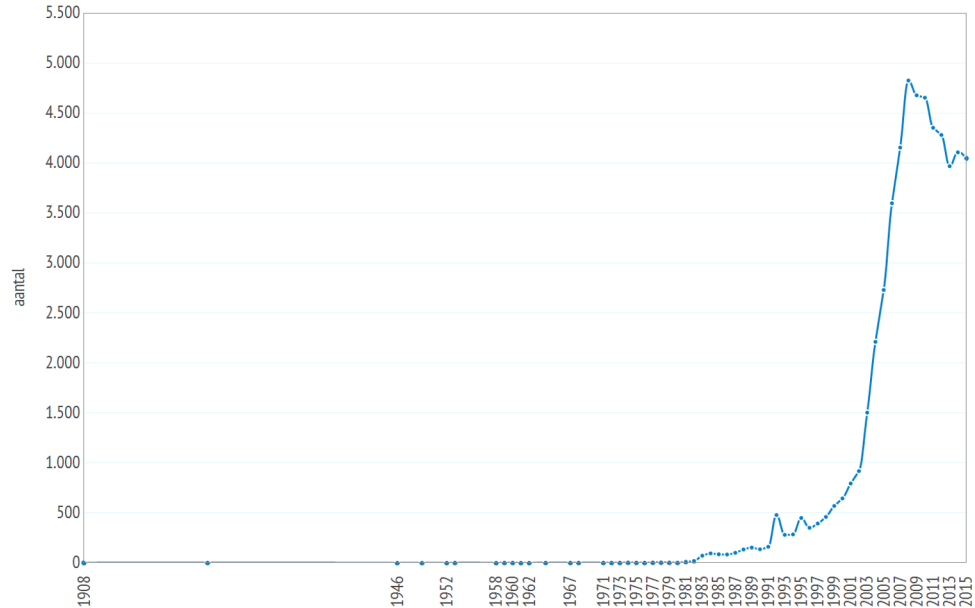


Alfie Talks-Liens
University of York
[@AlfieTalks](#)

Problem: too much text

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

Archeologische onderzoeksmeldingen in Nederland - 1908-2015



Eenheid: aantal

Bron: RCE - Monumentenregister - ARCHIS



AGNES v2.0 (beta)

Zoek door 60.000 archeologische rapporten uit het [DANS archief](#).

Let op: deze versie is nog niet uitgebreid getest, gebruik op eigen risico. Mocht zich een foutmelding voordoen, 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:

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.

Tijdperiode

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

Begin jaar:

Eind jaar:

Specifiek zoeken

Geen goede resultaten? Probeer te zoeken op concepten:

Artefact:

Context:

Soortnaam:

Niet exact zoeken

Gebruik 'fuzziness' 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.

Synonymy: Neolithic

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

Polysemy: Swifterbant

- Time period
- Excavation event
- Pottery type
- Place

Named Entity Recognition (NER)

... the excavation in **Swifterbant** was ...
... fragments of **pottery** nearby ...
... dated to the **Iron Age**. Other finds ...



NER

We found a stone axe from the Bronze Age

0 0 0 MAT ART 0 0 PER PER

Assign a label to each token (word)

Either nothing (0) or an entity type (Material, Artifact, Period, etc)

NER Methods

- Rule-based
- 'Classic' Machine Learning (CRF)
- Deep / Transfer Learning (BERT)
 - ▷ Generic model
 - ▷ Archaeology specific model

Rule-based NER



- **GATE software** (no labeled data, quick!)
- Vocabulary check
 - if 'axe' in vocab: mark as artifact
- Patterns / Rules
 - if [1 or 2 numbers] AND ends with 'th century': mark as time period

Conditional Random Fields NER

- 'Classic' machine learning
- Supervised learning, needs labelled data
- Like SVM, but takes context into account (prev/next word)
- Baseline for NER (although somewhat overtaken by spaCy)

Transfer Learning with BERT

(Bidirectional Encoder Representations from Transformers)

- Currently state of the art in NLP
- Pre-train neural network on large generic dataset
- Model 'learns' language (by trying to predict words)
- Fine-tune model on small labeled dataset to predict entities (or do other tasks)



BERT Models

- Google's multilingual model
- Language specific models (BERT-base, BERT-base-german, BERTje)
- Archaeology specific models
 - ▶ Take language model, further pre-train with (lots of) archaeology data
 - ▶ Thanks to data from ADS, DANS, Heidelberg uni journals

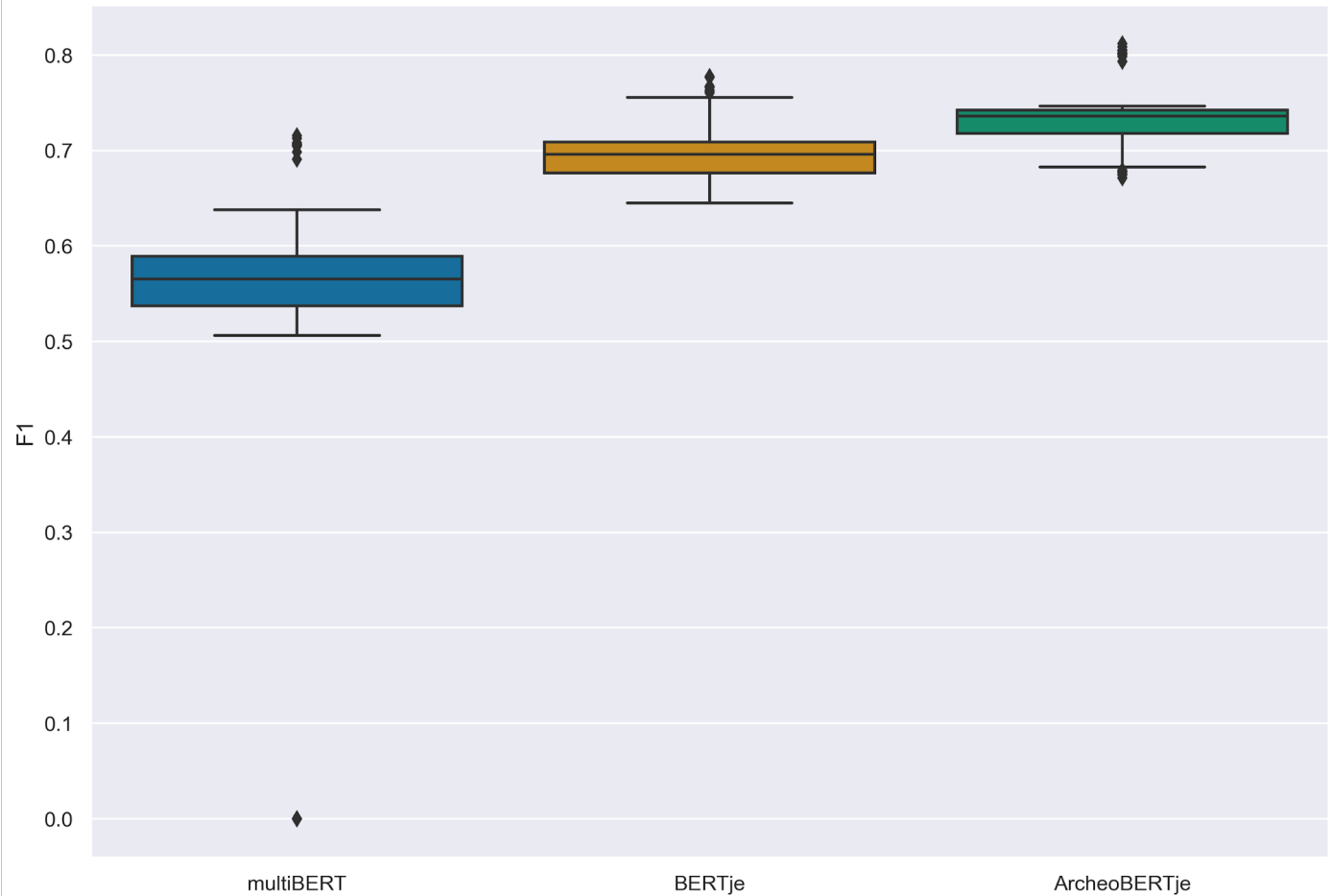


Results

5 fold cross validation, micro average, ART and PER only

Method	Language	Precision	Recall	F1
Rule based	NL	0.667	0.626	0.646
CRF	NL	0.773	0.628	0.690
General BERT	NL	0.772	0.721	0.746
Archaeo BERT	NL	0.789	0.761	0.774
Rule based	DE	0.410	0.537	0.465
CRF	DE	0.668	0.391	0.476
General BERT	DE	0.699	0.654	0.676
Archaeo BERT	DE	0.702	0.703	0.703
Rule based	EN	0.765	0.573	0.655
CRF	EN	0.794	0.670	0.721
General BERT	EN	0.794	0.785	0.789
Archaeo BERT	EN	0.720	0.585	0.646

Results (Dutch)



Labelling Full Dutch Corpus

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, <u>150 - 210</u> , late medieval, modern
Total	13,695,105	361,202	

's Gravenhage (The Hague)

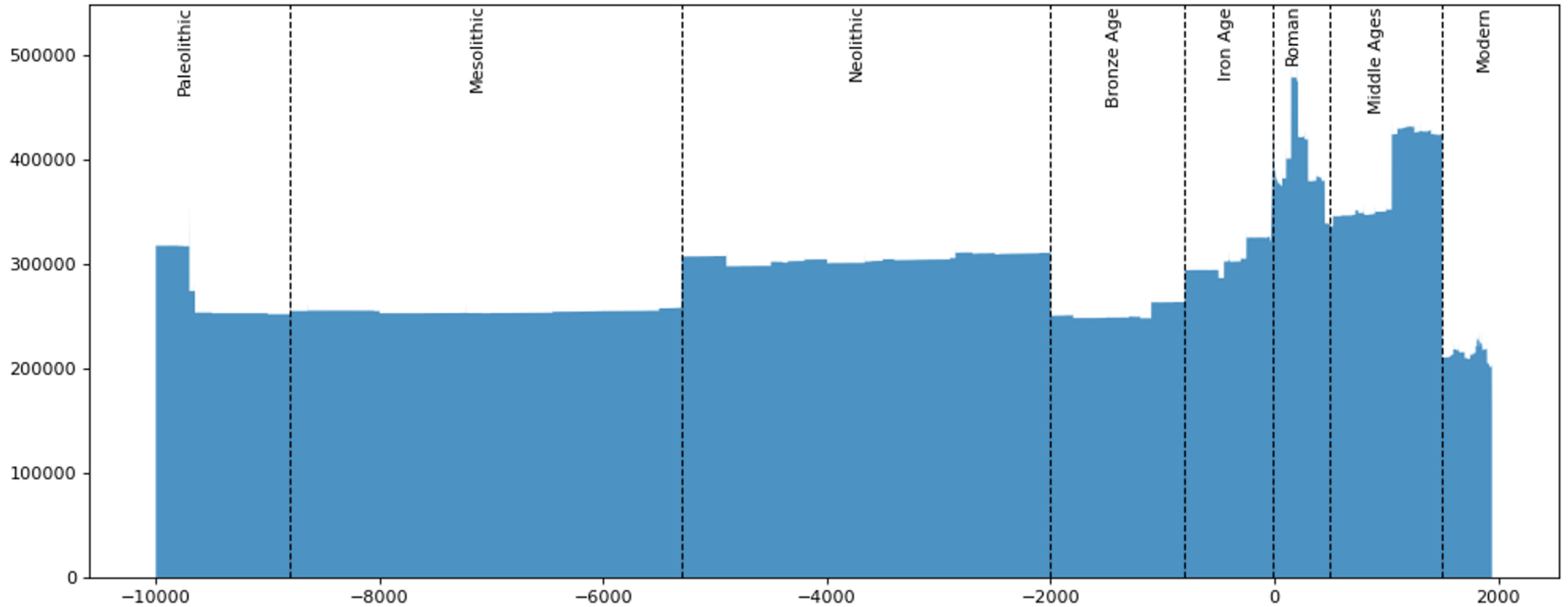


Extracting Year Ranges

- Rules based script / regexp
- 'Neolithic' → -5200 / -3000 (using thesaurus)
- 'second half 3rd century' → 250 / 300
- '1450 ± 50 BP' → 450 / 550
- 1250 BC → -1250 / -1250

See also new work by Binding, C., & Tudhope, D. (2023). Automatic Normalization of Temporal Expressions. *Journal of Computer Applications in Archaeology*, 6(1), 24–39. <https://doi.org/10.5334/JCAA.105>

Extracting Year Ranges



Conclusions

- Rule-based easy, no training data, not far off CRF, does need vocab
- BERT works best, but computationally expensive
- Creating archaeology specific BERT
 - ▶ Boosts performance
 - ▶ Less variance in output
- German is harder than Dutch and English (as it is for humans)



Thanks!

Any questions?

You can find me at:



@alex_brandsen



a.brandsen@arch.leidenuniv.nl

ChatGPT



AL

Pretend you are an archaeologist, annotating artefacts and time periods in a text string. Artefacts are defined as man-made objects found in excavations, and time periods are any temporal reference made in text. Now annotate the following sentence:



"We have found a bronze axe in a well, close to an intact Swifterbant pot from the early Neolithic"



Artefacts:



- Bronze axe

Time periods:

- Early Neolithic: This is a specific temporal reference provided in the sentence. The early Neolithic period is generally dated to around 4000-3000 BCE in Europe, which is a time before the Bronze Age.
- Swifterbant culture: The Swifterbant culture is a prehistoric culture that existed in the Netherlands and Belgium during the Late Mesolithic and Early Neolithic periods, roughly between 5600-3400 BCE. The intact Swifterbant pot found in close proximity to the bronze axe suggests that the axe may have been deposited in the well during the later stages of the Swifterbant culture, or in a later time period after the culture had ended.

ChatGPT



AL

Now annotate the following text, and output the results in the CONLL format:

"The pottery from the Medieval period includes Coarse border ware, Deritend ware and Humber ware. The Coarse border ware has been dated to 13th to 14th centuries AD"

< 2 / 2 >



Text annotation in CONLL format

1 The O
2 pottery O
3 from O
4 the O
5 Medieval B-TIME_PERIOD
6 period I-TIME_PERIOD
7 includes O
8 Coarse B-ARTEFACT_TYPE
9 border I-ARTEFACT_TYPE
10 ware I-ARTEFACT_TYPE
11 , O
12 Deritend B-ARTEFACT_TYPE
13 ware I-ARTEFACT_TYPE
14 and O
15 Humber B-ARTEFACT_TYPE
16 ware I-ARTEFACT_TYPE
17 . O

ChatGPT



AL

Now annotate this sentence, which is also about archaeology, but in Dutch:

"Uit de 7e eeuw v. Chr. zijn 2 messen gevonden van vuursteen, waarschijnlijk toe te kennen aan de Swifterbant cultuur"



Text annotation in CONLL format

1 Uit O
2 de O
3 7e B-TIME_PERIOD
4 eeuw I-TIME_PERIOD
5 v. I-TIME_REFERENCE
6 Chr. I-TIME_REFERENCE
7 zijn O
8 2 B-ARTEFACT_COUNT
9 messen I-ARTEFACT_TYPE
10 gevonden O
11 van O
12 vuursteen I-MATERIAL
13 , O
14 waarschijnlijk O
15 toe O
16 te O

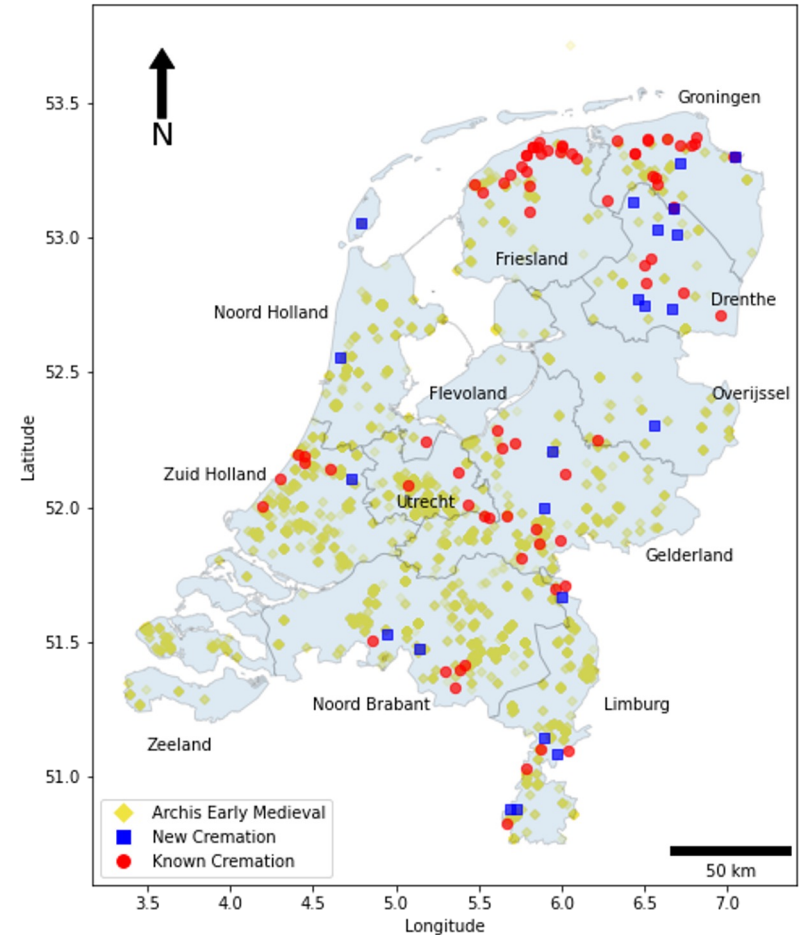
AGNES v2.0

- Archaeological Grey-literature Named Entity Search
- ArcheoBERTje NER, indexed in ElasticSearch
- 60,000 documents
- agnesearch.nl



Evaluation

- Focus group: overall positive
- Case study on Early Medieval cremations:
 - ▷ Found 30% more than previously known
 - ▷ Found sites in unexpected areas
 - ▷ Changing our views on Early Medieval burial practices



Future work: EXALT

- Expanding and improving AGNES, 4 year project
- Adding more languages (English, German) ✓
- Adding more reports (UK, Belgium, Germany)
- Adding more document types (papers, books)
- End goal: multilingual semantic search for Dutch archaeology

ArcheoBERTje Errors (on training data)

		O	B-ART	I-ART	B-MAT	I-MAT	B-PER	I-PER	B-CON	I-CON	B-LOC	I-LOC	B-SPE	I-SPE	
True label	O	391053	1598	351	145	0	700	665	728	17	575	228	368	122	
	B-ART	1721	6492	265	183	1	38	0	136	0	15	0	136	0	
	I-ART	717	363	1086	22	5	3	6	1	1	0	2	26	4	
	B-MAT	197	372	60	559	5	0	0	2	0	10	0	19	1	
	I-MAT	23	4	32	9	10	0	0	0	0	2	0	1	0	
	B-PER	1086	14	4	0	0	6997	232	4	1	17	0	5	0	
	I-PER	1169	2	1	0	0	267	5917	0	0	0	0	0	1	
	B-CON	1627	182	3	0	0	46	0	3415	24	5	0	0	0	
	I-CON	93	1	3	0	0	0	0	45	24	0	2	0	0	
	B-LOC	788	10	0	2	0	21	1	3	0	3545	61	5	0	
	I-LOC	446	3	0	0	0	0	0	0	0	98	638	0	0	
	B-SPE	389	168	7	31	0	8	0	8	0	0	0	2186	44	
	I-SPE	124	8	53	1	4	0	0	0	0	0	0	40	542	
			O	B-ART	I-ART	B-MAT	I-MAT	B-PER	I-PER	B-CON	I-CON	B-LOC	I-LOC	B-SPE	I-SPE
			ArcheoBERTje predicted label												