Peer Community In Archaeology

Excavating archaeological knowledge with Text Mining, NER and BERT

Daniel Carvalho *based on peer reviews by* ***Simon Carrignon*** *and 1 anonymous reviewer*

Lasse van den Dikkenberg, Alex Brandsen (2025) Using Text Mining to Search for Neolithic Vlaardingen Culture Sites in the Rhine-Meuse-Scheldt Delta. Zenodo, ver. 2, peer-reviewed and recommended by Peer Community in Archaeology.
https://doi.org/10.5281/zenodo.14763691

Submitted: 09 August 2024, Recommended: 10 February 2025

Cite this recommendation as:
Carvalho, D. (2025) Excavating archaeological knowledge with Text Mining, NER and BERT. *Peer Community in Archaeology*, 100547. 10.24072/pci.archaeo.100547

Published: 10 February 2025

The production of texts in archaeology is vast and multiple in nature, and the archaeologist often misses the true extent of its scope. Machine learning and deep learning have a top place to play in these analyses (Bellat et al 2025), with text extraction methods being therefore a useful tool for reducing complexity and, more specifically, for uncovering elements that may be lost in the midst of so much literary production. This is what Van den Dikkenberg and Brandsen set out to do in the specific case of Vlaardingen Culture (3400-2500 BCE). By using NER (Named Entity Recognition) with BERT (Bidirectional Encoder Representations from Transformers) they were able to recover data related to the location of sites, the relevance of the data and, just as importantly, potential errors and failures in interpretation (Van den Dikkenberg and Brandsen 2025). The contextual aspect is emphasized here by the authors, and is one of the main reasons why BERT is used, which is logically a wake-up call for the future: it is not enough to classify or represent data, it is essential to understand what surrounds it, its contexts and its particularities (Brandsen et al 2022).

For this, refinement is always advocated, as these models need constant attention in terms of both training data and parameters. This constant search means that this article is not simply an analysis, but that it can be a relevant contribution both to the culture in question and to the way in which we approach and extract relevant information about the grey literature that archaeology produces. Thus, Van den Dikkenberg and Brandsen present us with an article that is eminently practical but which considers the theoretical implications of this automation of the search for the contexts of archaeological data, which reinforces its relevance and, consequently, its recommendation.

References:

Bellat, M., Orellana Figueroa, J. D., Reeves, J. S., Taghizadeh-Mehrjardi, R., Tennie, C. & Scholten, T. (2025). Machine learning applications in archaeological practices: A review. https://doi.org/10.48550/arXiv.2501.03840

Brandsen, A., Verberne, S., Lambers, K. & Wansleeben, M. (2022). Can BERT dig it? Named entity recognition for information retrieval in the archaeology domain. Journal on Computing and Cultural Heritage, 15(3), 1–18. https://doi.org/10.1145/3497842

Van den Dikkenberg, L. & Brandsen, A. (2025). Using Text Mining to Search for Neolithic Vlaardingen Culture Sites in the Rhine-Meuse-Scheldt Delta. Zenodo. v2 peer-reviewed and recommended by Peer Community In Archaeology https://doi.org/10.5281/zenodo.14763691

Reviews

Evaluation round #1

DOI or URL of the preprint: https://doi.org/10.5281/zenodo.13283972
Version of the preprint: 1

Authors' reply, 29 January 2025

Download author's reply

Decision by Daniel Carvalho, posted 22 December 2024, validated 24 December 2024

Dear all,

Please see the important recommendations of both reviewers, especially regarding the exposition of the model you intend to show. Some other aspects related to data visualization seem equally pertinent to make the best use of your narrative.

Reviewed by Simon Carrignon, 24 September 2024

The paper propose an interesting approach that uses large language model to detect archaeological sites from a specific culture that that may have been missed in previously build database using archaeological reports. The method rely on searching for specific terms in publicly available pdfs, and use of LLMs to ensure the hits are well charactised and not false positives.

ALthough I do thinkg the paper is interesting and present an original and well caried piece of work, there is still a few points I would like the authors to answer.

First of all, it is still not very clear to me the relationship between AGNES, NER, BERT, and the relevance classifications presented in Tables 2 to 4. What is BERT bringing exactly? When given the search terms in Table 1 and thanks to its training, is it able to classify reports that fall within "Vlaardingen Groep," (for example) even if the terms are not necessarily mentioned or not in this exact order or spelling in the report? Then, how are the reports classified as relevant or not? Is that done by BERT too, or is it manually checked?

This is very likely due to my lack of knowledge of the model used, and of the previous papers cited by the authors, but I am struggling to understand how this is different from simply looking for the term (with some regex) and then manually check the reports returned. I do trust the authors that this is indeed different, but I struggle to see how exactly this is done, and I think the authors could clarify this with one or two sentences.

The web interface of Agnes is also briefly mentionned, and when digging the supplementary material, one can find the IP address to interact with the logiciel. And again, this may be due to my own ignorance about the

project and previous publication, but why no address to access AGNESS is given yet? is it because the project is still in developpement? Also, is the code behind AGNES, and the interactions with BERT and the NER methods going to be published? The overal architecture of the project seems to be an amazing achievement, that could be used by any country/institutoin with pulicly available archaeological reports in pdf, is there any plan to publish/share these tools? I understand they may not be ready but maybe the authors could mentionned it in the paper.

I am not too sure about the network visualization of the results. They may be a nice alternative to simple barplot with the percentage per query, but then the authors should say a sentence or two justifying this choice and explaining a bit more what do these network tells us.

To follow with these network: what is the difference between multiple edges (like between ”Vlaardingen Stein” and ”not relevant”) vs thicker edges (like between ”Vlaardingen Stein Wartberg” and ”semi-relevant”)? Also, wouldn't some normalization of the edge/node sizes help visualize the data? such as taking the log or the square root of the values?

Finally: I very much appreciated the discussions about the discrepancy between the results of this paper and the known literature, and how the authors interprate it. This also make a good case on how AI can help understand how history and sociology of a discipline can bias th results. I found the maps particularly interesting in that regards, especially the one that shows how the culture is described in different reports, revealing non-random customs among the archaeologists excavating these sites. Nonetheless, I think these maps could be better rendered. The choice of color and the large size of the dots used to represent the sites make it very hard to see clear patterns. Do the geology maps really need to be included? maybe only some DEM would be enough? Many of the colors between the dots and the base map are very similar, making it difficult to interpret.

Using smaller dots to represent the sites, on a map with fewer colors —perhaps even including only rivers and the sea, if slopes still makes it too messy?- could allow to remove the 'zoomed' versions of the maps? Then the two figures could be joined in a single one with two panels: previously found/not found on one side and cultural attribution on the other? Pushing it further: as there are only four categories for found/not found, different shapes of circles could be used to encode this, allowing to have everything on one single combined map? However, I understand this might make it less readable. Maybe the different shape could then be combined with the year of the excavation, that may illustrate nicely the fact that older report are not in pdf and show how the missing sites are the one published long time ago

Minor remark:

- What does ”OG”, in the sentence ” the OG Large Language Model (Devlin et al. 2019)” means? as it's not defined before i read it as Original Gangster, but I doupt the authors wanted to use this temrinology?
- The output of the csv supplemtary material of the table 2 as like 1000 row but only 167 sites have something, is that

Overall this is an interesting paper, that showcase a nice use of AI in archaeology, and I think it would fit well PCI Archaeology. Modulo the comment I make here I think it will be a great publication!

Reviewed by anonymous reviewer 1, 25 November 2024

Disclaimer I am a geographer and archaeologist trained in new archaeological methods, and I have worked on machine-learning applications in archaeology for a few years. However, I am neither a Dutch archaeologist specialist nor a computer scientist. Therefore, my knowledge of Dutch archaeological cultural problematics and large language model hyper-tuning is limited. I am also a non-native English speaker.**Title and abstract:**
Major issue:

- *Does the title clearly reflect the content of the article? [ ] Yes, [x] No (please explain), [ ] I don't know* "From the River to the see" sentence in the title does not fit any of the elements discussed in this pre-print. If I agree with using a "pre-title" or a short sentence that might attract the reader here, it does not seem to correlate with the rest of the subject. One major problem is using the Large Language Model word to describe BERT (Cf. below).

- *Does the abstract present the main findings of the study? [ ] Yes, [x] No (please explain), [ ] I don't know.* The abstract is clear for most of the points except for the additional goal of the collected results besides testing the model. In the discussion, several points are debated on the VLC that are not mentioned in the abstract. I would recommend adding one or two short sentences on these cultural interpretations.

Minor issue:

- **Line 18** There is a lack of clarity in describing the Vlaardingen Culture characteristics.
- **Lines 38 – 39** The keywords use words similar to those in the title. Alternative ones for a broader range of results in ulterior referencing online would be suggested.

Introduction: Major issue:

- *Are the research questions/hypotheses/predictions clearly presented? [] Yes, [x] No (please explain), [ ] I don't know.* As a non-expert in Dutch archaeology, my comment might be inaccurate. However, I would suggest a deeper explanation of the VLC culture, which is described here as "The vast majority of Vlaardingen Culture sites consist of artefact scatters without clear house plans (Van Gijn & Bakker 2005). Only a few sites contain clearly discernible house plans (Stokkel 2017; Van Beek 1990; Van Kampen 2013; Van Zoolingen 2021; Verhart 1992)." (**l.63-66**).

- *Does the introduction build on relevant research in the field? [x] Yes, [ ] No (please explain), [ ] I don't know*

Structure:

The introduction's structure would need major revision, as the description of the objective (**lines 59 – 60**) comes before the description of the VLC and AGNES. This paragraph of lines **59 – 70** would need to be set at the end of the introduction or rewritten entirely so it does not mention elements that have not been introduced before.

For the whole 1.2 part (**lines 111 – 124**) I would suggest instead of the paragraph a table synthetizing all the information and including **lines 121 – 124** in the method part or legend of the table.

I would appreciate a more detailed paragraph on the recent influence of large language models in general and the explosion of transformers after Vaswani et al. 2017 paper.

Language:

The English is good, but it is not easy to read. It would win in being read by a native English speaker:

- **Lines 47 – 49** The sentence is unclear, especially the "the end users" meaning.
- **Line 56**: Present first the full name of the tool and then the acronym in parenthesis, not the other way around.
- **Lines 72 – 77** The paragraph on Brandsen and Lippok's previous study, 2021, is unclear as it repeats itself.
- **Line 116** A general comment: All the prompts should be written in their native language (Dutch) in italics and then in English, like "*Archeologie*/Archaeology."

Content:

On the general content of the introduction, one primary piece of information is missing: a clear description of the Vaardingen Culture. Only one sentence "The vast majority of Vlaardingen Culture sites consist of artefact scatters without clear house plans" (l.63-65) is giving information on this culture. However, a proper description of a chrono-cultural entity should include more information. How are we differentiating this culture from the other at the same chronological time (e.g. ceramic, metalwork, buildings)? What are its specificities?Minor issue:

- **Line 63** There is no reference for the VLC chrono-cultural period.
- **Lines 96 – 99** No map is provided for representing this new area of excavation, or the "potential" extent of the VLC.
- **Lines 116, 118-119** It would be needed to provide a reference or hyperlink for the mentioned database/portals.

Materials and methods: Major issue:

- *Are the methods and analyses sufficiently detailed to allow replication by other researchers? [ ] Yes, [x] No (please explain), [ ] I don't know.* The main goal of this preprint is to provide a detailed process of the AGNES tool. However, if the reference to the Brandsen 2023 BERT model is present, no detailed information is furnished on the preprocessing of the data set or the model.

- *Are the methods and statistical analyses appropriate and well described? [ ] Yes, [x] No (please explain), [ ] I don't know.* There seems to be confusion about the Large Language Model definition.

Structure:

The general structure of the chapter is clear except for the 2.2.1 part, which stands alone without any reason. Maybe it would need to include paragraph **lines 202 – 208** as part of the 2.2 part only. As it mentions the relevant table, it should be placed above **line 190.**

Language:

Similar general considerations as for the Introduction part stand here (Cf. above):

- **Lines 132 – 133** This sentence seems unnecessary.
- **Lines 136 – 139** The sentence is unclear and needs re-writing.
- **Line 137** The spelling of "medieaval" while correct for British writing only, is not commonly used. Please consider writing "medieval" without an "a" if submitting to an American journal afterwards.
- **Lines 137 – 138** With the same considerations as in the previous part, please include the prompts' original language (Dutch) in italics.
- **Line 146** The same consideration is given to the prompt and the "mediaeval" writing.
- **Line 168** Table 1 The English translation does not need capital letters.
- **Line 183** Unclear sentence.

Content:

Enhanced the choice of BERT as in Brandsen 2023 "However, for specific domains and low-resource languages, BERT can still outperform LLMs."

A critical point is the confusion between LLMs and the BERT model, as expressed in " The NER is done using BERT (Bidirectional Encoder Representations from Transformers), the OG Large Language Model (Devlin et

al. 2019). Similar to the newer GPT (Generative Pre-trained Transformer) models, BERT uses large amounts of unlabelled text data to pre-train a model, gaining an understanding of words and their contexts" (**l.149 – 152**). I do not know if the authors refer to an "OG" model, which would be an LLM, in which case you will need to give more precisions and a reference, or if they refer to BERT as an LLM. This would be inaccurate in the second case as the BERT does not have two main characteristics of LLMs. First, the number of parameters of BERT is "only" hundreds of millions, while the number of GPT3 or PALM parameters is hundreds of billions. Second, the BERT does not have a decoder phase in its architecture, contrary to GPT3 or other LLMs (Rogers et al. 2020). This confusion would need to be clarified as many parts of the pre-print mention BERT as an LLM would need to be reformulated.

The details of the pre-process are not given, which is another major issue. The paper would need to refer to the pre-training of the BERT model, both original and modified versions trained on the 60k documents of the DANS (Brandsen, 2023). Information on which BERT model is used (large or base) would be needed. The pre-treatment of a text (e.g. deleting of white space, common words, numbers) needs to be transparent to be replicable, and no information is given here.Minor issue:

- **Lines 165 – 166** There is no reference for the VLC chrono-cultural period.
- **Lines 168** Table 1 would need to have combined rows for similar English translations.
- **Lines 190 – 192** If categories 10 – 12 are mentioned, the mention of categories 1 – 9, related to a previous typology from Brandsen and Lippok, 2021, would be needed.

Results:

Major issue:

- *In the case of negative results, is there a statistical power analysis (or an adequate Bayesian analysis or equivalence testing)? [x] Yes, [ ] No (please explain), [ ] I don't know*
- *Are the results described and interpreted correctly? [x] Yes, [ ] No (please explain), [ ] I don't know*

Structure:

No comment on the structure of this part

Content:

The results part fit every requirement.

Language:

Similar general considerations as for the Introduction part stand here (Cf. above):

- **Lines 215, 216, 218, 223, 235, 241-244, 268** All the prompts should be written in italics as in a foreign language (Dutch).
- **Line 210** In English writing, the separator for decimal numbers is a dot "." and not a comma "," (9.7%).
- **Line 284** Figure 3 The legend "not in AGNES" is not clear to the reader.

Minor issue:

- **Lines 210 – 211** The numbers should be kept, and the per cent written under parenthesis 439 relevant hits (9.7%; Table 4)
- **Lines 220, 238, 272** Table 4, Table 5 Table 6 Add a column for the percentages.
- **Lines 231-234, 248-250** Figure 1, Figure 2 Add, if possible, the percentage of each node.
- **Line 259** Add the percentage for the 89 sites under parentheses.

- **Line 284** Figure 3 Why choose a palaeogeographical map when the environment is not discussed in the discussion part? The coordinates and projection systems are required on the left map.

Discussion and conclusion: Major issue:

- *Have the authors appropriately emphasized the strengths and limitations of their study/theory/methods/argument? [ ] Yes, [x] No (please explain), [ ] I don't know.* I would have liked a more detailed discussion on the implication of such "by-catch" techniques for further research in grey literature (**l.379-395**). Only a few lines describe future possibilities, while there are many, and can also overcome strong bias, such as the absence of specialists in some excavations, which led to underestimating or misinterpreting some scattered findings. With this approach, this bias could be counterbalanced.
- *Are the conclusions adequately supported by the results (without overstating the implications of the findings)? [x] Yes, [ ] No (please explain), [ ] I don't know*

Structure:

The general structure of the chapter is clear except for **lines 309-313**, which would fit better in the introduction as they define the different cultures. **Lines 357 – 361** would also refer to the results and not the discussion part.

In the conclusion, the paragraph from **lines 444 – 449** does not entirely fit into the conclusion and would need to be either rewritten or changed into the discussion part.

Content:

There is one general lack of development of the discussion. As a non-specialist of the VLC, I cannot provide information on whether the newly founded site would improve our knowledge of this culture. However, the interpretation and possible uses of "by-catch" are limited to a few lines (**l. 390 -395**), while its possibilities extend to many areas and timelines and could help fix bias from the survey, in particular when specialists are missing.

Another comment on the F1 score (**l.397 – 398**) is whether it would be possible to recall the already identified sites (Found previously and in AGNES = 39).

Language:

Similar general considerations as for the Introduction part stand here (Cf. above):

- **Lines 398, 401** In English writing the separator for decimal numbers is a dot "." and not a comma ",".
- **Lines 426-428** An unclear sentence and a repetition of the word "aimed".
- **Line 428** The word "aimed" is used once again.
- **Lines 460 – 461** Redundant sentence.

Minor issue:

- **Line 329** Figure 4The same comments are made as for figure 3.
- **Line 418** A reference or hyperlink would be needed for *PyMUPDF* software.
- **Line 431** Thirty and its percentage.
- **Line 459** The number before the percentage.

Literature:

The choice of literature seems quite reasonable overall. I would only suggest recent publications on the uses of LLMs in archaeology (Agapiou and Lysandrou, 2023; Cobb, 2023; Lapp and Lapp, 2024), and more especially the book of Gonzalez-Perez et al. (2023) Discourse and Argumentation in Archaeology: Conceptual and Computational Approaches with several chapters on NLP or text extraction. **Conclusion:**

In conclusion, as such, I cannot recommend this paper for publishing. It needs major revisions. The problems reside in the confusion of the large language model, the lack of context on the Vlaardingen Culture, and the methodology workflow needing to be more transparent.

Download the review