Final Report:
Finding Fossils in the Literature

In partnership with the Neotoma Paleoecology Database.

Ty Andrews Jenit Jain Kelly Wu Shaun Hutchinson

2023-06-28

Executive Summary

The project Finding Fossils in Literature is sponsored by the Neotoma database (Neotoma) which houses a paleoecology database. The challenges Neotoma faces are 1) researchers have to manually enter sample data into Neotoma, 2) researchers are not aware of Neotoma or that their research fits into it, and 3) there are too many articles published for the Neotoma team to monitor new research. This project has 3 primary deliverables to solve the challenges, first is an article relevance prediction model which predicts whether newly published articles are relevant to Neotoma. Second, is an article data extraction pipeline which identifies key entities such as taxa or geographic location. Last is a data review tool for Neotoma data stewards to review the extracted data before it is submitted to Neotoma. Once the extracted data has been reviewed, the corrections will be used to retrain the first two models of the pipeline.

Table of contents

1 Introduction

The Neotoma database (Neotoma) (Williams et al. 2018) serves as a valuable resource for researchers investigating ecological transformations spanning the last 5 million years. Nevertheless, the collection of data heavily relies on manual submissions from researchers, presenting challenges in data entry and impeding collaborative endeavors. This project aims to automate the extraction of data from relevant journal articles which can be added to Neotoma. This is completed in three parts. First, relevancy to the Neotoma database is predicted. Relevant articles are then parsed using natural language processing techniques. Finally, a data review tool is used to review and correct the extracted data before it is submitted to Neotoma. The outputs of this data review tool are then used to improve the two models in the future.

2 Data Science Methods

2.1 Article Relevance Prediction

The article relevance prediction was posed as a binary classification problem using article metadata to predict whether the article is relevant to Neotoma.

2.1.1 Approach

2.1.1.1 Building the Training Data

To train the supervised classification model, a sample of labelled articles was compiled as shown in Table 1.

Table 1: Labelled Sampled Article

Label	Sample Size	Description
Positive	911	A randomly sampled list of articles that currently contribute to Neotoma was provided as the positive cases. These articles cover various types of fossil data in the Neotoma database.
Negative	3523	Articles from non-paleoecology-related subjects were queried to form a representative sample of non-relevant articles.

The labelled articles were split into train/validation/test sets with splits of 70%/15%/15%, which correspond to 3103/665/666 articles respectively.

2.1.1.2 Preprocessing & Feature Selection

The article's metadata were obtained from the CrossRef API (Crossref 2023), which provided over 50 types of metadata. To reduce dimensionality, feature selection was conducted with results in Table 2.

Table 2: Feature Selection Decisions

Feature	Decision
Title & subtitle	Title and subtitle (if any) are concatenated as a descriptive text feature. Text representation techniques were then applied to this feature.

Feature	Decision
Abstract	Less than 50% of articles have abstracts available due to copyright issues. To address this, the available abstracts are combined with the article title and subtitle to create a descriptive text feature. A binary feature indicates whether the article has an abstract or not, which can affect model predictions.
Author & Journal	In order to mitigate potential bias towards well-established authors and journals, the feature list was modified to exclude time-sensitive indicators to account for new Authors and Journals published in the future.
Subject	The subject of the journal was used as a proxy of the name of the journal.
Number of citation	Number of citations could indicate if the article contains important data and is referred to in other articles.
Feature related to the publication process	These features do not provide any contextual information about the article, thus were removed from the feature list.

The features selected include a descriptive text feature (concatenation of title, subtitle and abstract), subject of the journal, number of citations and a binary indicator for having an abstract available from CrossRef or not.

2.1.1.3 Feature Engineering: Text representation

The descriptive text feature provides crucial contextual information for the model to predict if the topic was related to paleoecology or not. Feature engineering was conducted to represent this descriptive text feature in numeric form. The following text representations were explored and tested to identify the most effective approach.

Bag of words representation (Baseline): word-frequency based method. This achieved moderate results but is limited by not containing contextual information.

Sentence embedding (Final): Sentence embedding is the state-of-the-art approach for text representation. Three sentence embedding models were experimented and allenai/specter2 model resulted in the best overall performance as documented in Table 3.

Table 3: Sentence Embedding Models

Model	Result
bert-tiny-finetuned-squadv2	This model was of a small scale so that the embedding process would be quick.
Biobert	This model was pre-trained based on large-scale biomedical corpora, thus it could be better at understanding biology-related jargons that frequently appear in the field of paleoecology.
specter2	This model was pre-trained using over 6 million scientific articles, thus it could be better at understanding the language style and pattern in scientific articles.

2.1.1.4 Model Selection

The following selection of Probability Classifiers, Analogy-based and Tree-based & Gradient-boosted models were experimented with.

Probability classifiers: Traditional probability classifiers provide fast training, high interpretability, and prediction time. Logistic regression and naive Bayes were experimented during model selection.

Analogy-based models: The article relevance prediction problem can be framed as looking for articles that are similar to the positive examples. K-nearest neighbour and Support Vector Machine with RBF kernel were experimented.

Tree-based & Gradient-boosted models: Tree-based classifiers use different feature subsets during the prediction, which can be suitable for high dimensional data problem. Decision Tree, Random Forest Model, LightGBM, XGBoost and CatBoost were experimented.

The chosen model went through further hyperparameter tuning to optimise the performance.

2.1.2 Model Evaluation

The main goal of article relevance prediction is to identify as many relevant articles as possible so that valuable research data can be discovered. The cost of false negatives will be high because missed articles are unlikely to be discovered by the Neotoma community where as false positives are lower cost with easy rejection in the data review tool. For this reason recall was optimized for.

Model interpretability was also important because the data likely has biases due to positive examples coming from existing articles in Neotoma. A more interpretable model would help identify model biases and make corrections in the long run.

2.2 Article Data Extraction

The article data extraction was posed as a Named Entity Recognition (NER) problem. NER is done by using models to predict the correct label for each word in the text.

2.2.1 Data Description

An original sample of 300+ full text articles related to Neotoma was provided. Some articles were non-english and were removed from the dataset. The data of interest to be extracted and thus labelled in the articles using NER were as follows:

- **SITE**: the specific site name given to the items studied
- **REGION**: general geographic regions around the site to give context to sites
- **GEOG**: geographic coordinates of sites or samples
- **TAXA**: the taxa uncovered and researched in the article
- **AGE**: any ages relevant to the samples
- **ALTI**: the altitude at which samples were studied
- **EMAIL**: any researchers emails for further follow up

2.2.2 Data Preprocessing and Labelling

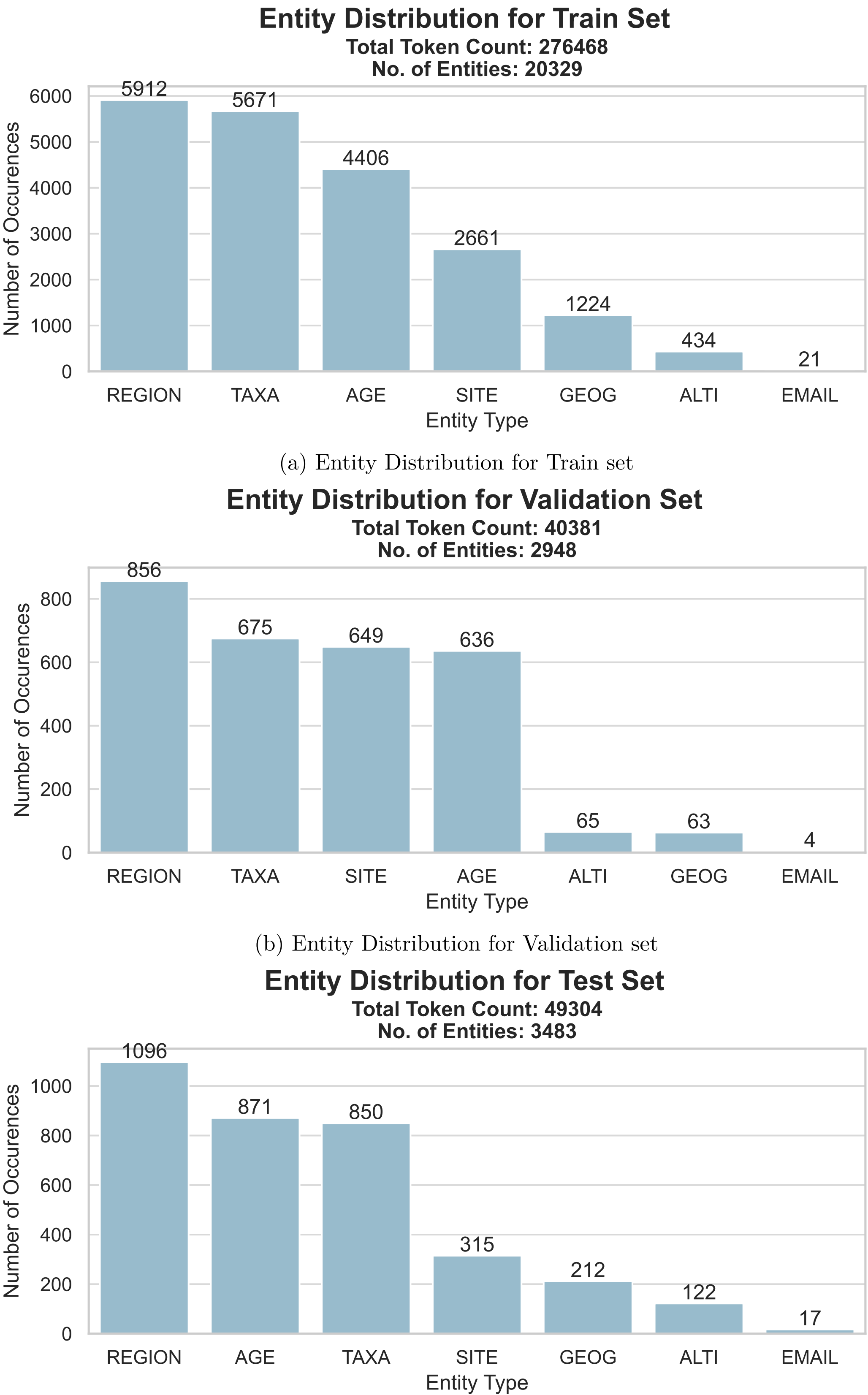
For the development of the NER models, a sample of 39 articles relevant to Neotoma were preprocessed and labelled by the research team using LabelStudio("Label Studio: Data Labeling Software" 2023). Other text preprocessing techniques such as lemmatization were not used due to text quality issues from OCR making it difficult to identify the correct lemma. The SITE entities were difficcult to identify and label due to obscure location names or an acronym of a core sample being used. Similarly, correctly identifying where one TAXA entity ends and the next begins proved challenging.

The labelled articles were split into train/validation/test sets by full article with splits of 70%/15%/15% respectively. The data splitting was done based upon whole articles to prevent data leakage into the model learning certain articles patterns before evaluation. This resulted in the entity distributions between train/validation/test as shown in Figure 1.

2.2.3 Approach

NER tasks are dominated by transformer based models (Popel and Bojar 2018). The best performing non-transformer based models are transition-based stack long-short term memory (S-LSTM) models used in the spaCy package ("spaCy NER" 2023).

The approaches considered along with the rationale for their inclusion/rejection from development are outlined in Table 4.



Entity Distribution for Train Set
Total Token Count: 276468
No. of Entities: 20329

(a) Entity Distribution for Train set

Entity Distribution for Validation Set
Total Token Count: 40381
No. of Entities: 2948

(b) Entity Distribution for Validation set

Entity Distribution for Test Set
Total Token Count: 49304
No. of Entities: 3483

(c) Entity Distribution for Test set

Figure 1: Entity Distribution

Table 4: NER Approaches

Approach	Rationale
Rule Based Models	This served as the baseline using regex to extract known entities but was not developed further due to the known issues with text quality due to OCR issues and infeasibility for entities like SITE.
Transition Based Model	Low computational cost, ideal for low-compute resource deployment
BERT Based Models	Many state of the art apporaches use BERT, multiple base models pre-trained on different domains are available
Text to Text Transfer Transformer (T5)	T5 models make each problem text to text based, for NER this requires significant pre/post processing to work and is excluded for this reason.

For the transformer based models two approaches were used for training, spaCy command line interface (CLI) ("spaCy NER" 2023) and HuggingFace's Training application programming interface (API) ("HuggingFace" 2023). Each have advantages and disadvantages which are outlined in Table 5.

Table 5: spaCy CLI andHugging Face's Training API Advantages and Disadvantages

	Pro	Con
spaCy Config Training	• Can integrate with any transformer hosted on HuggingFace • Prebuilt config scripts that require minimal changes • Access to spaCy's unique transition-based NER model • Ability to integrate multiple ML models in an ensemble	• Knowledge of bash scripting required • Limited configuration options
Hugging Face Trainer API	• Able to use non-NER models and add classification head easily • Able to easily integrate custom tracking of metrics and external logging to MLflow	• More code required

Using the Hugging Face training API multiple models were trained and evaluated. Each base model along with the hypothesis behind it's selection is outlined in Table 6.

Table 6: Hugging Face Model Hypotheses

Model	**Hypothesis**
RoBERTa-base	One of the typically best performing models for NER (Wang 2020)
RoBERTa-large	A larger model than the base version with potential to learn more complex relationships with the downside of larger compute times.
BERT-multilanguage	The known OCR issues and scientific nature of the text may mean the larger vocabulary of this multi-language model may deal with issues better.
XLM-RoBERTa-base	Another cross language model (XLM) but using the RoBERTa base architecture and pre-training.
Specter2	This model is BERT based and finetuned on 6M+ scientific articles with it's own scientific vocabulary making it well suited to analyzing research articles.

Final hyper parameters used to train the models are outlined in Table 7.

Table 7: Hugging Face Model Training Hyperparameters

Parameters	**Notes**
Batch size	- Maximized to utilize all available GPU memory, 8 for RoBERTa based models
Gradient Accumulation	- Used to mimic larger batch sizes, this value was set at 4 to achieve batch sizes of ~12k tokens based on best practices (Popel and Bojar 2018)
Epochs	- Initial runs with 10-20 epochs, observed evaluation loss minima occurring in first 2-8, settled on 10
Learning Rate	- Initially 5e-5 was used and observed rapid over fitting with eval loss reaching a minimum around 2-4 epochs then increasing for the next 5-10 - Moved to 2e-5 as well as introducing gradient accumulation of 3 epochs to increase effective batch size

Parameters	Notes
Learning Rate Scheduler	- All training was done with a linear learning rate scheduler which linearly decreases learning rate across epochs
Warmup Ratio	- How many steps of training to increase LR from 0 to LR, shown to improve with Adam optimizer - (AI 2023) Set to 10% initially

Using the spaCy CLI, the two models were trained and evaluated with each models advantages and disadvanctages outlined in Table 8.

Table 8: spaCy CLI Model Advantages and Disadvantages

Model	Advantages	Disadvantages
RoBERTa-base	• State-of-the-art pretrained transformer for NLP tasks in English • Context rich embeddings	• Computationally expensive to train and inference • Cannot fine-tune • Slow processing
en_core_web_md	• A smaller word vector model with static embeddings for words • Memory and speed efficient	• Static embeddings without context • Cannot handle OOV words well • Cannot fine-tune

Final hyper parameters used to train the spaCy models along with comments on each impact are outlined in Table 9.

Table 9: spaCy CLI Final Hyperparameters

Parameters	Notes
Batch size	- Maximized to utilize all available GPU memory, 128 for transformer based model and 512 for word vector based model
Epochs	- Initial runs with 15 epochs, observed evaluation loss minima occurring in first 7-13 depending on learning rate
Learning Rate	- Initial learning rate of 5e-5

Parameters	Notes
Learning Rate Scheduler	- Warmup for 250 steps followed by a linear learning rate scheduler
Regularization	- L2 (lambda = 0.01) with weight decay
Optimizer	- Adam (beta1 = 0.9, beta2=0.999)
Early stopping	- 1600 steps

Both the spaCy and HuggingFace models used RoBERTa as the base model. The spaCy model labels a sequence of inputs using an algorithm similar to transition-based dependency parsing using (Stack-LSTMs)(Guillaume Lample 2016). Whereas the HuggingFace transformer model added a linear layer with softmax on top of the RoBERTa embeddings to perform the token classification.

The workflow in Figure 2 shows the primary steps for training with the Entity Extraction models with intermediate files and processes.

2.2.4 Model Evaluation

The target for the data extraction is to maximize recall. This was chosen as it maximizes the amount of entities extracted and minimizes the amount of time a reviewer needs to read the full article during review. Entity and token based recall are used as the primary metrics for model evaluation. Entity based recall checks all tokens of an entity matching are correct. Token based recall allows for partial matches of entities which contain multiple words. Higher token based recall is preferred as it maximizes data extracted and utilizes the data review process to correct any false positives.

2.3 Data Review Tool

To facilitate the review and improve the efficiency of Neotoma data stewards, an interactive dashboard was developed as the final data product. This dashboard enables manual review of the Article Relevance Prediction and Article Data Extraction results. Users can compare extracted entities to sentences or access the full-text articles to make corrections. They also have the ability to delete incorrect entities and add any missed entities. The Data Review Tool generates a parquet object as output, which can be used to retrain the Article Entity Extraction model and update the Neotoma database. This process promotes enhanced information sharing and improved results, reducing the time required for data stewards to review extracted entities in articles.

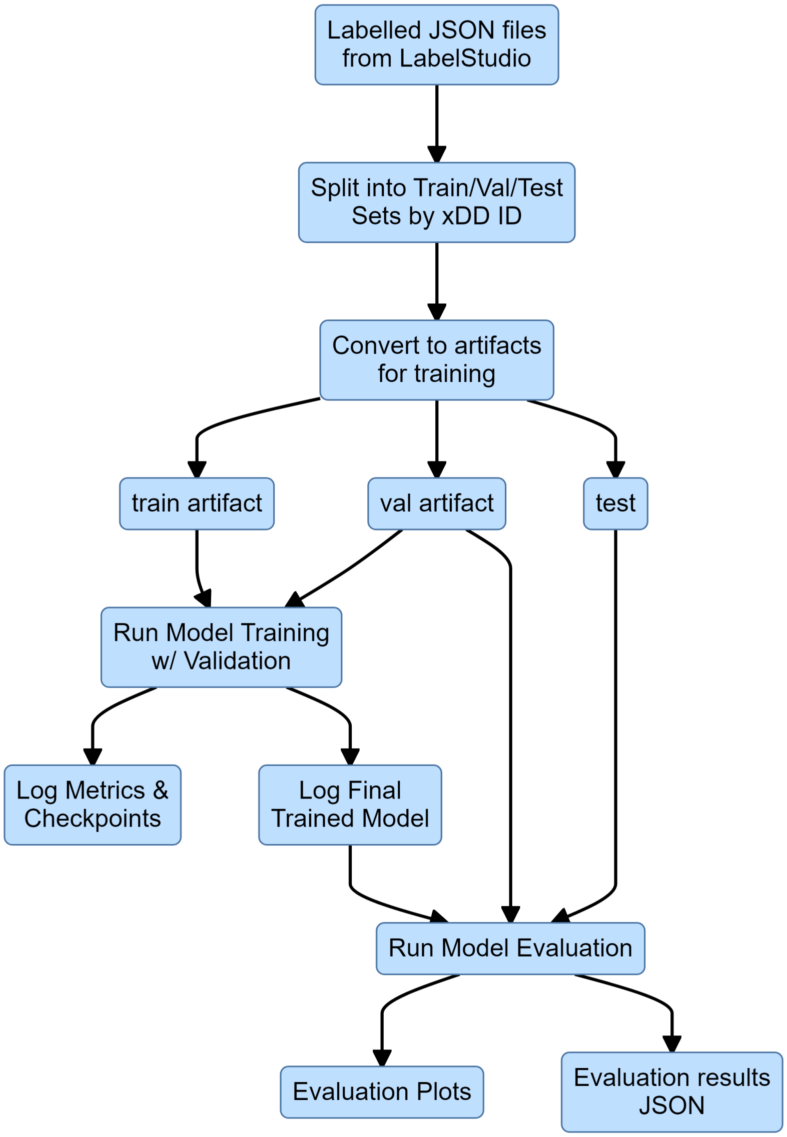

Labelled JSON files from LabelStudio
Split into Train/Val/Test Sets by xDD ID
Convert to artifacts for training
train artifact
val artifact
test
Run Model Training w/ Validation
Log Metrics & Checkpoints
Log Final Trained Model
Run Model Evaluation
Evaluation Plots
Evaluation results JSON

Figure 2: The Entity Extraction model training process with intermediate files and processes.

2.3.1 Approach

The project's objective was to develop a customizable and freely available product with custom components. Considering this, the decision was made to prioritize open-source solutions such as R Shiny (Chang et al. 2023) and Plotly Dash ("Dash" 2023) over paid alternatives like Tableau or Power BI. The choice between R Shiny and Plotly Dash was influenced by the project's customization requirements. During discussions with Neotoma, the decision to implement either of these options did not have any impact. Ultimately, the team concluded that Plotly Dash in Python would be the preferred choice for the ongoing development of the project dashboard as the rest of the pipeline was to be written in Python.

2.3.2 Evaluation

In order to create an interactive tool that would be appropriate and efficient for the reviewers who are using this tool, the aim was to make it user-friendly and simple to use. To accomplish this success criteria, the requirements outlined in Table 10 were developed along with the targets to make these requirements successful for the data reviewers.

Table 10: Data Review Tool Target Metrics

Requirement	**Target**
Options for reviewing extracted data	Accept, Reject, Edit then Accept
Other data made available to the user	Article DOI, Journal Name, Hyperlink to Article
Displaying text from where the data was extracted	Current sentence and 1-2 sentences before/after.
User skill to run	Non-Technical (e.g. no code/CLI)
Number of mouse clicks to review single piece of data	1-2
Reviewing workflow	Able to save/resume progress.
Output file format	Parquet

3 Data Products and Results

3.1 Article Relevance Prediction

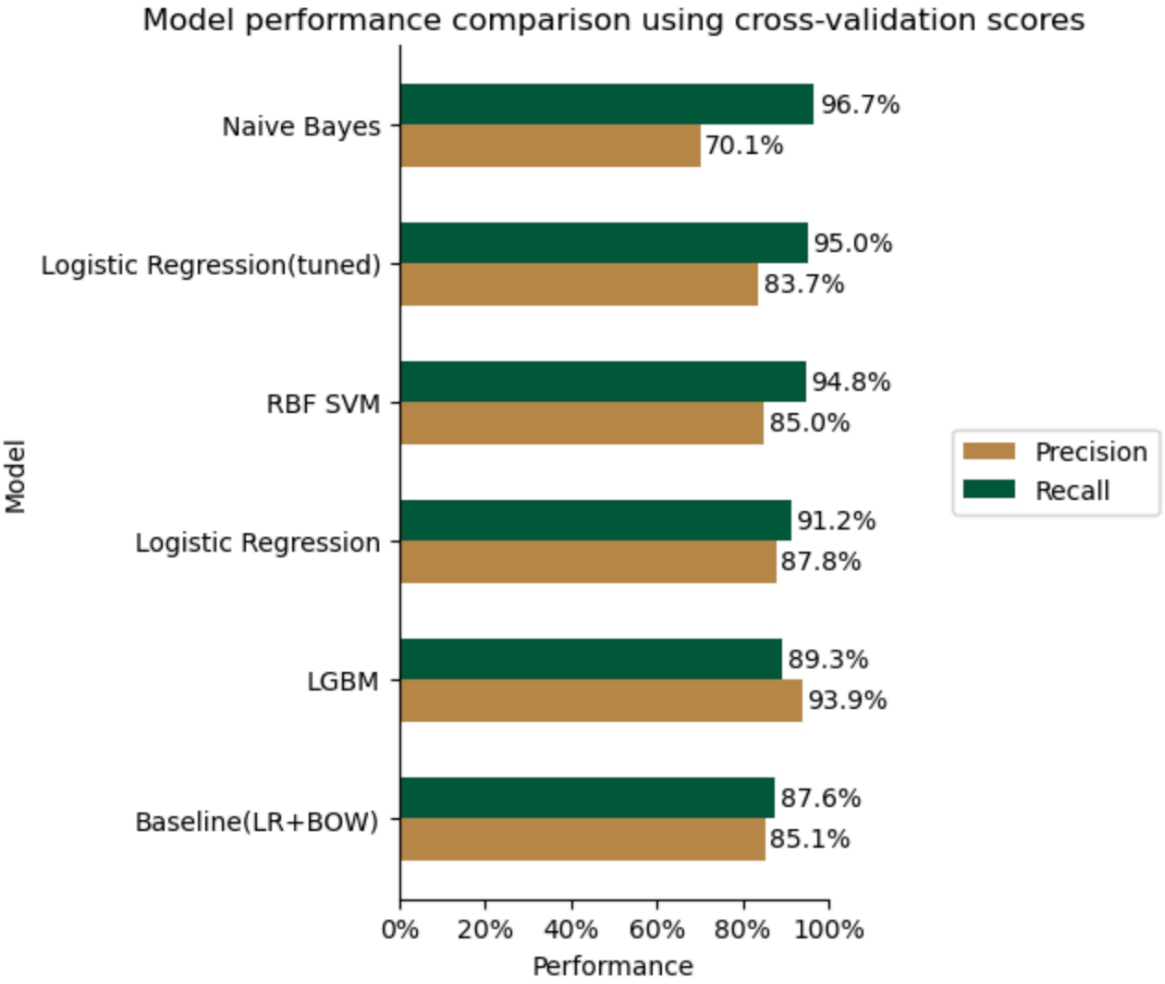


Figure 3: Model Performance Comparison using Cross-Validation Scores

Among all the models trained, Naive Bayes had the highest recall score (96.7%), but its precision score was low (70.1%). Among all gradient boosted models, LGBM had the highest precision (93.9%), but its recall score (89.3%) did not outperform logistic regression (94.8%). Among the analogy-based models, SVM with RBF kernel achieved a recall performance (94.8%) that's comparable to logistic regression (95.0%), and both had sufficiently high precision (Figure 3). The final decision was to use logistic regression with tuned hyperparameters for its high performance and better interpretability. On the test data set, the final model achieved a recall score of 96.5%, and a precision score of 85.2%. To put this into actual article counts, among the 666 articles in the test set, there are 5 false negatives and 24 false positives.

3.2 Article Data Extraction

From training spaCy and Hugging Face models there are two candidate models proposed. The first is the Hugging Face finetuned RoBERTa based model which performed best of the Hugging Face models. The second is spaCy's transformer based model. The rationale for two models moving forward is the spaCy transformer model had better entity based recall along with better precision, meaning the extracted data was more often correctly extracted. Whereas the Hugging Face RoBERTa model achieves better token recall but has lower precision, meaning it detects more of the overall entities but contains more false detections which must be deleted by the reviewers. The differences in the recall and precision can be attributed to the different NER models leveraging the RoBERTa-base transformer embeddings differently.

Figure 4 shows the primary candidate models token and entity based recall.

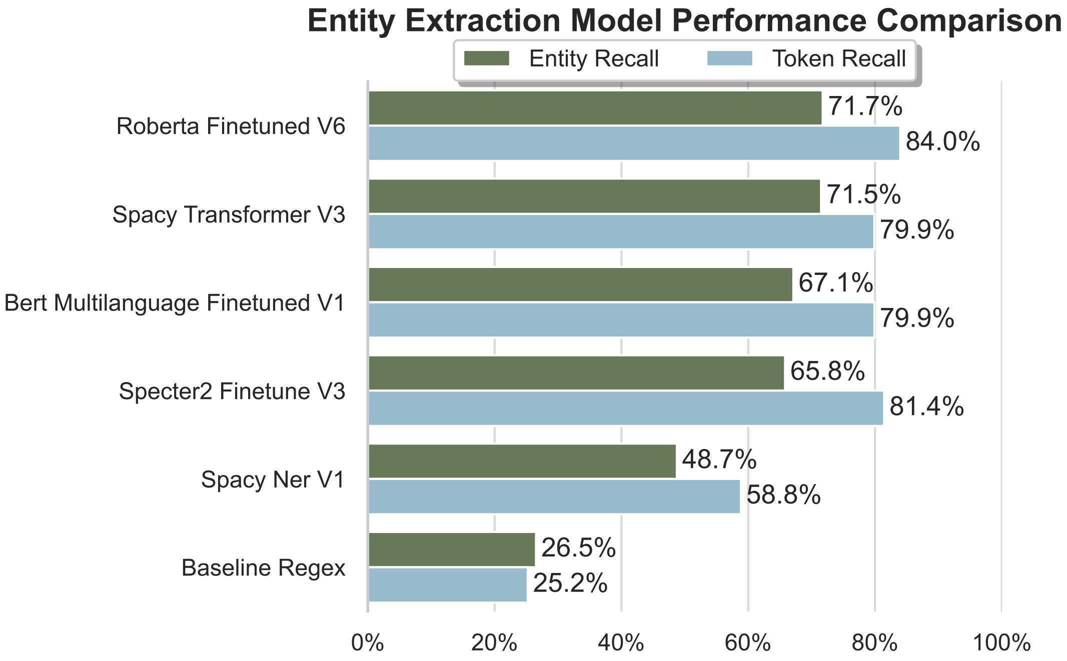


Figure 4: The entity and token based recall for each model on the test set.

The token based recall for each model on the test set by entity type are shown in Table 11.

Table 11: Entity extraction model results for token based recall, overall and by entity type

Model	Recall (%)	Age (%)	Site (%)	Taxa (%)	Region (%)	Geog (%)	Email (%)	Alti (%)
RoBERTa	84	89.3	67.6	91.4	86.9	44.8	100	76.2
Specter2	81.4	86.7	83	86.9	71.8	62.1	47.6	53.1
BERT Multi-lang.	79.9	81.1	76.9	85.5	74.9	58.6	33.3	81.5
SpaCy Transf.	79.9	82.2	73.7	88	70.6	64.7	100	83.3
SpaCy NER	58.8	63.6	23	69.7	66.2	35.3	0	26.2
Baseline Regex	25.2	9	0	46.3	26.9	0	100	27.2

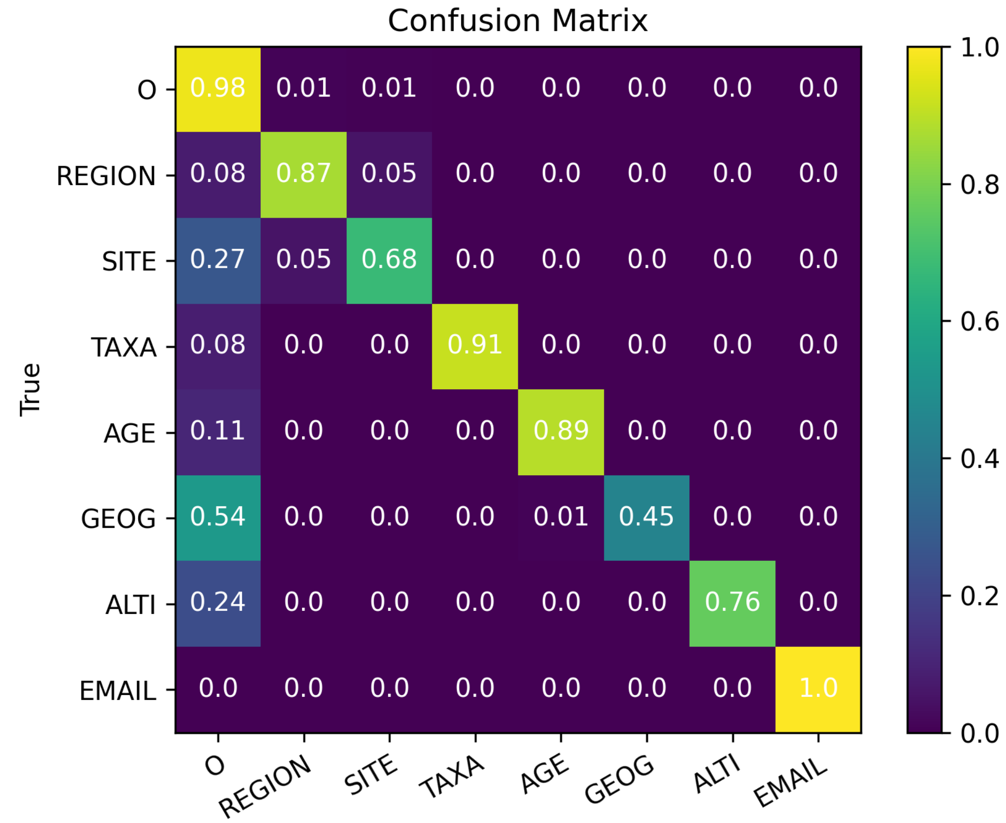
An important observation to make here is that the top models had a lower precision score for the SITE names and REGION names. The models got confused when deciding whether an entity should be classified as a SITE or a REGION. This was partially due to quality of labeling entities as well as the fact that both these types correspond to the name of a place or a wider area. See Figure 5 for a confusion matrix generated using the test set assets highlights the issue.

3.3 Data Review Tool

The final data review tool that was created is a multi-page Plotly Dash ("Dash" 2023) application. The tool can be replicated by launching Docker ("Docker" 2023) containers, enabling anyone within the Neotoma community to easily utilize the tool for reviewing outputs from the pipeline.

This web application enables users to review the extracted entities from articles, make changes, add additional missed entities and remove incorrectly extracted entities. The entire review process does not require any coding knowledge, and reviewers can navigate through the review workflow in three clicks or fewer from the Article Review page Figure 6.

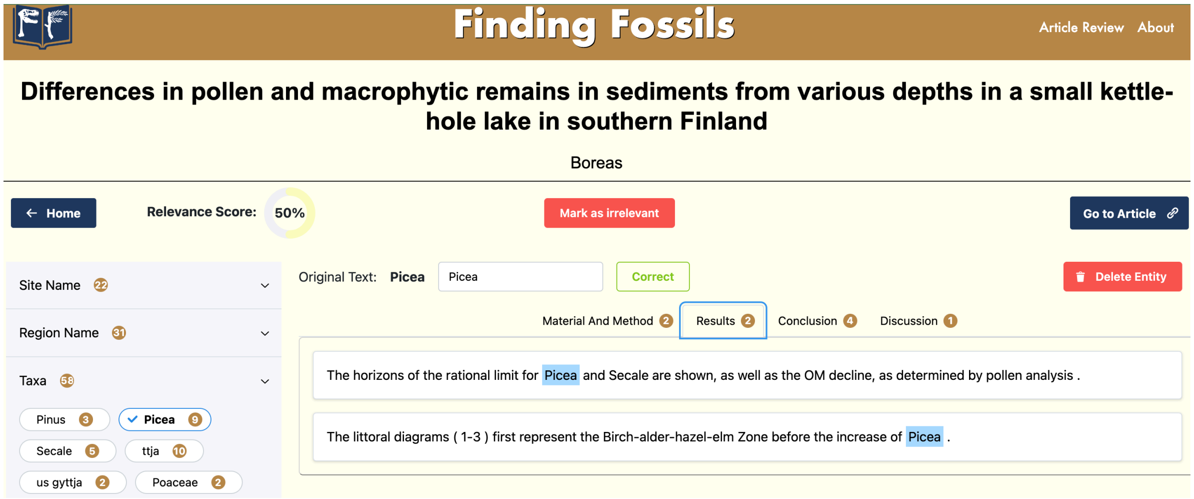
The output of this data review tool is a parquet file that stores the originally extracted entities as well as the corrected entities. The Neotoma organization can utilize these entities to add new paleoecology articles to the database. In addition, the reviewed entities can be fed back to model using our model retraining pipeline, contributing to the continuous improvement of the quality of extraction of entities as outlined in Table 12.



Confusion Matrix

True	O	REGION	SITE	TAXA	AGE	GEOG	ALTI	EMAIL
O	0.98	0.01	0.01	0.0	0.0	0.0	0.0	0.0
REGION	0.08	0.87	0.05	0.0	0.0	0.0	0.0	0.0
SITE	0.27	0.05	0.68	0.0	0.0	0.0	0.0	0.0
TAXA	0.08	0.0	0.0	0.91	0.0	0.0	0.0	0.0
AGE	0.11	0.0	0.0	0.0	0.89	0.0	0.0	0.0
GEOG	0.54	0.0	0.0	0.0	0.01	0.45	0.0	0.0
ALTI	0.24	0.0	0.0	0.0	0.0	0.0	0.76	0.0
EMAIL	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0

Figure 5: Confusion Matrix for RoBERTa Model



Differences in pollen and macrophytic remains in sediments from various depths in a small kettle-hole lake in southern Finland

Boreas

Relevance Score: 50%

Site Name 22

Region Name 31

Taxa 58

Pinus 3
Picea 9
Secale 5
ttja 10
us gyttja 2
Poaceae 2

Original Text: **Picea** Picea

Material And Method 2 | Results 2 | Conclusion 4 | Discussion 1

The horizons of the rational limit for Picea and Secale are shown, as well as the OM decline, as determined by pollen analysis .

The littoral diagrams (1-3) first represent the Birch-alder-hazel-elm Zone before the increase of Picea .

Figure 6: Data Review Tool

Table 12: Data Review Tool Metric Results

Requirement	Results
Options for reviewing extracted data	Accept, Reject, Edit, Add then Accept
Other data made available to the user	Article DOI, Hyperlink to Article
Displaying text from where the data was extracted	Current sentence and 1 sentence before/after.
User skill to run	Non-Technical (e.g. no code/CLI)
Number of mouse clicks to review a single piece of data	3 clicks from launch
Reviewing workflow	Able to save/resume progress.
Output file format	Parquet

3.4 Product Deployment

The end goal of this project is to have each data product running unsupervised. The article relevance prediction pipeline was containerized using Docker ("Docker" 2023). It is expected to run on a daily or a weekly basis by Neotoma to run the article relevance prediction and submit relevant articles to xDD (Peters, S.E., I.A. Ross, T. Rekatsinas, M. Livny 2021) to have their full text processed.

The Article Data Extraction pipeline is containerized using Docker and contains the entity extraction model within it. It will be run on the xDD servers as xDD is not legally allowed to send full text articles off their servers. The container accepts full text articles, extracts the entities, and outputs a single JSON object for each article. The JSON objects are combined with the article relevance prediction results and loaded into the Data Review Tool. Figure 7 depicts the work flow.

4 Conclusion and Recommendations

4.1 Conclusions

The Neotoma database plays a crucial role in paleoecological research by providing data on ecological changes over the past 5 million years. However, the reliance on manual submissions

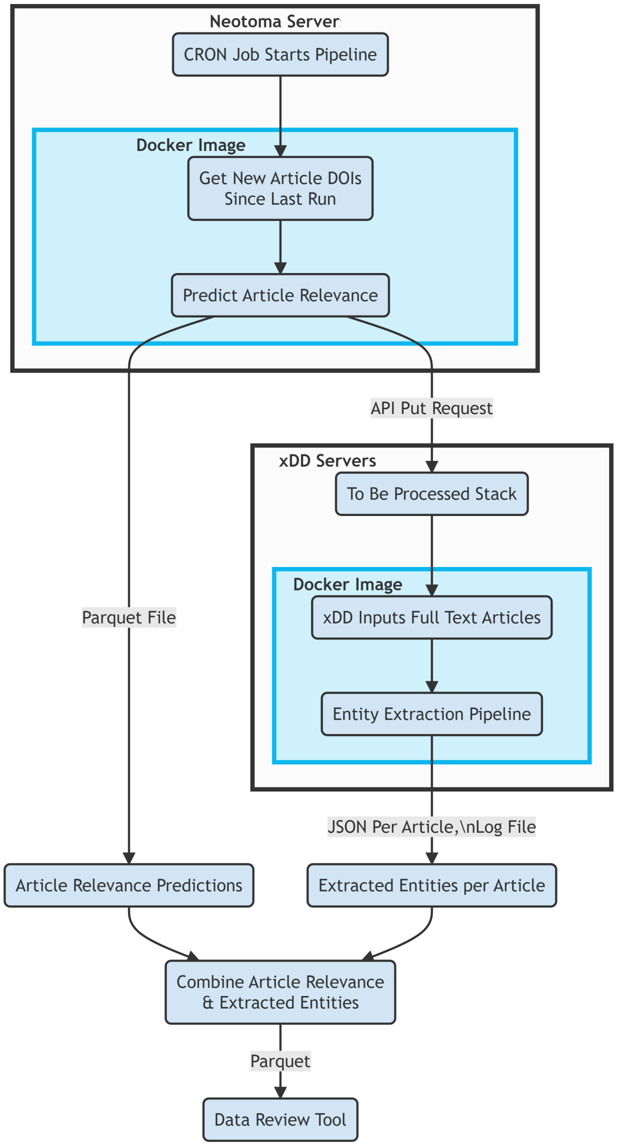

Neotoma Server
CRON Job Starts Pipeline
Docker Image
Get New Article DOIs Since Last Run
Predict Article Relevance
API Put Request
xDD Servers
To Be Processed Stack
Docker Image
xDD Inputs Full Text Articles
Entity Extraction Pipeline
Parquet File
JSON Per Article,\nLog File
Article Relevance Predictions
Extracted Entities per Article
Combine Article Relevance & Extracted Entities
Parquet
Data Review Tool

Figure 7: How the MetaExtractor pipeline flows between the different components.

by researchers has led to difficulties in data entry and possibly hindered collaborative efforts to fully understand these ecological changes. The implementation of our pipeline brings significant benefits to the Neotoma community and paleoecological research. Firstly, by automating the search for relevant articles through the Article Relevance Prediction this process will grow the Neotoma community, attracting more researchers and fostering collaboration in ecological research. Secondly, researchers will experience reduced time and effort required for data uploads, as some of the manual submission process is replaced by the Article Entity Extraction and Data Review Tool. Finally, with the addition of more data, researchers in the Neotoma community may be able to answer questions about ecological changes that were previously unknown.

4.2 Recommendations

Due to time constraints, improvements were identified but could not be implemented within the project's timeframe. Three essential features have been identified that would enhance the overall pipeline.

1. **Relevance Review Page:** incorporating an article review page would allow for a thorough assessment of article relevance, reducing both false positives and false negatives in the predictions. It will also reduce the bias in the training dataset that currently exists due to our data collection strategy.
2. **Adding new entities:** There are additional entities that were considered beneficial but not mandatory by the Neotoma team. Adding additional entities can be done in LabelStudio and could benefit the Neotoma community at large.
3. **Correction Model** There are likely many corrections that will be performed frequently on the data review tool. Therefore, a model could be implemented to learn these corrections and preprocess these corrections to reduce time spent reviewing.

5 Acknowledgements

Data were obtained from the Neotoma Paleoecology Database (http://www.neotomadb.org) and its constituent database(s). The work of data contributors, data stewards, and the Neotoma community is gratefully acknowledged.

A huge thanks to Simon Goring & Socorro Dominguez from the Neotoma Database team for their support on this project.

References

AI, Borealis. 2023. "Tutorial 17: Transformers III - Training." https://www.borealisai.com/research-blogs/tutorial-17-transformers-iii-training/.

Chang, Winston, Joe Cheng, JJ Allaire, Carson Sievert, Barret Schloerke, Yihui Xie, Jeff Allen, Jonathan McPherson, Alan Dipert, and Barbara Borges. 2023. *Shiny: Web Application Framework for r*. https://shiny.rstudio.com/.

Crossref. 2023. "Crossref REST API." https://www.crossref.org/services/metadata-delivery/rest-api/.

"Dash." 2023. Plotly. https://dash.plotly.com.

"Docker." 2023. Docker. https://www.docker.com/.

Guillaume Lample, Sandeep Subramanian, Miguel Ballesteros. 2016. "Neural Architectures for Named Entity Recognition." *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, March. https://doi.org/http://dx.doi.org/10.18653/v1/N16-1030.

"HuggingFace." 2023. https://huggingface.co/.

"Label Studio: Data Labeling Software." 2023. https://github.com/heartexlabs/label-studio.

Peters, S.E., I.A. Ross, T. Rekatsinas, M. Livny. 2021. "xDD API." JSON. geodeepdive.org.

Popel, Martin, and Ondrej Bojar. 2018. "Training Tips for the Transformer Model." *CoRR* abs/1804.00247. http://arxiv.org/abs/1804.00247.

"spaCy NER." 2023. Explosion. https://spacy.io/api/entityrecognizer.

Wang, Yu. 2020. "Application of Pre-Training Models in Named Entity Recognition." *2020 12th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*. https://doi.org/10.1109/IHMSC49165.2020.00013.

Williams, J. W., E. G. Grimm, J. Blois, D. F. Charles, E. Davis, S. J. Goring, R. Graham, et al. 2018. "The Neotoma Paleoecology Database: A Multi-Proxy, International Community-Curated Data Resource." *Quaternary Research* 89: 156–77.