The changing landscape of text mining - a review of approaches for ecology and evolution

Maxwell J. Farrell[1-3*], Nicolas Le Guillarme[4], Liam Brierley[5], Bronwen Hunter[6], Daan Scheepens[7], Anna Willoughby[8], Andrew Yates[9], Nicole Mideo[1]

1. Department of Ecology & Evolutionary Biology, University of Toronto, Toronto, Ontario, Canada
2. School of Biodiversity, One Health & Veterinary Medicine, University of Glasgow, Glasgow, UK
3. MRC-University of Glasgow Centre for Virus Research, Glasgow, UK
4. Univ. Grenoble Alpes, Univ. Savoie Mont Blanc, CNRS, LECA, Laboratoire d'Ecologie Alpine, Grenoble, France
5. Department of Health Data Science, University of Liverpool, Liverpool, United Kingdom
6. School of Life Sciences, University of Sussex, Brighton, United Kingdom
7. Division of Biosciences, University College London, United Kingdom
8. Odum School of Ecology, University of Georgia, USA
9. Informatics Institute, University of Amsterdam, Amsterdam, Netherlands

* Corresponding author: maxwell.farrell@glasgow.ac.uk

Abstract

In ecology and evolutionary biology, synthesis and modelling of data from published literature is a common practice for generating insight and testing theories across systems. However, the tasks of searching, screening, and extracting data from literature are often arduous. Researchers may manually process hundreds to thousands of articles for systematic reviews, meta-analyses, and compiling synthetic datasets. As relevant articles expand to tens or hundreds of thousands, computer-based approaches can increase efficiency and dramatically improve the transparency and reproducibility of literature-based research. Methods available for text mining are rapidly changing due to developments in machine learning-based language models. Here we review the growing landscape of approaches, mapping them onto three broad paradigms (Frequency-based approaches, Traditional Natural Language Processing, and Deep learning-based language models). This review serves as an entry point to learn foundational and cutting edge concepts, vocabularies, and methods, and foster better integration of these tools into ecological and evolutionary research. We discuss approaches for modelling ecological texts, generating training data, developing custom models, and interacting with Large Language Models, and we present challenges and possible solutions to implementing these methods in ecology and evolution.

An overview of approaches

Tools from linguistics, computer science, and machine learning can aid in all phases of text-based research, such as identifying relevant papers, analysing research trends, constructing or expanding databases, and building automated pipelines to translate text into data ready for statistical analysis. Despite the power and utility of text mining, it has yet to be adopted widely in ecology and evolution [1]. Farrell et al. (2022) present recent uses in ecology and evolution, and discuss future applications, but a major gap remains for life science scientists to learn about available methods and their underlying philosophies. Here, we review the growing landscape of approaches and methodological paradigms for text mining. To build a vocabulary necessary for collaboration with linguists and computer scientists, throughout the review we define key terms and concepts (see Table 1 for definitions of underlined terms). Our aim is to encourage further uptake of text mining by detailing how individual steps can be tailored to goals in ecology and evolution and combined within an overall pipeline akin to other analytical fields [2]. We begin by describing the evolution of text mining tools and map them to three broad "paradigms": 1) Frequency-based Approaches, 2) Traditional Natural Language Processing, and 3) Deep Learning. As recent advances in NLP have turned to deep learning-based models, we highlight approaches to train models specific to ecology and evolution, make the most of the rich training data in existing ecological databases, and practices for interacting with language models for extraction and synthesis of data in ecology and evolutionary biology.

Paradigm 1 - Quantifying Text: Frequency-based Approaches

One of the most common outputs of text analysis is the word cloud, a hallmark of literature reviews, scientific talks, and personal websites of academics. This method takes a segment of text and identifies individual words, counting how often each appears, and displaying them with font size proportional to frequency. A word cloud is an example of wider "bag of words" approaches, a paradigm based on quantifying frequencies of words within and/or across documents. Here the meaning and order of words is ignored, instead treating them as discrete units, much like quantifying the abundances of species in a given location.

When considering words or phrases (collectively, terms) independent of their context, such as in a bag-of-words analysis, "stop words" are often filtered out. Stop words are common, uninformative words that provide no specific meaning or context (e.g. "the", "are", "we", "why"). Lists of stop words are included in NLP software libraries [3,4], but there is no single definitive list and choice of words to remove, or whether to filter words at all will depend on the task and nature of the texts to be analysed. Rather than filtering specific words, researchers may opt for a threshold approach by removing X% of the most and least common terms across documents. These choices reinforce that bag-of-words approaches do not capture the meanings of words, their order, or their contexts. In linguistics, word sense refers to the multiple meanings a word can take depending on its context. For example, "bark" could refer to the outermost layer of a woody plant, or the sound dogs make. Without context, a bag-of-words approach would consider these terms equivalent. Rather than using single words as terms of interest, further context can be captured in a frequency-based approach by focusing on groups of n words at a time in shifting windows, termed n-grams, (e.g. the phrase "text mining" is a 2-gram). How one chooses to go from raw text to distinct terms will change their number and frequency, ultimately influencing downstream analyses and interpretation.

Beyond word clouds, exploratory analyses of single terms or n-grams commonly involve quantifying the similarity of documents through document-term matrices (akin to site-by-species matrices in community ecology). Terms can be weighted by how frequently they occur in any document, for example by using a TF-IDF weighting (term-frequency-inverse document frequency), which takes into account the frequency of terms within and between documents. This captures the importance of a term to a document, adjusted for the fact that some words appear more frequently across the whole set of documents. TD-IDF matrices can be used for many downstream tasks including text summarization, keyword identification, and as a way to transform text and use it as a predictor or an input to downstream models. For example, 'Topic Modelling' is an unsupervised approach that uses a method like Latent Dirichlet Allocation to group documents into abstract topics that can be further explored. Nunez-Mir et al. (2016) review the ecological applications of topic models and other similar approaches to identify hidden themes in a body of literature and demonstrate their utility for systematic and exploratory literature reviews. Bag-of-words approaches can also be useful in text classification. For example, instead of unsupervised methods like Topic Modelling, labels can be attached to documents and used to develop models for information retrieval (e.g. systems that identify “relevant” documents). For example, Naive Bayes classifiers based on word frequencies can achieve >90% accuracy in some scenarios, such as identifying advertisements relevant to the wildlife trade [6].

Paradigm 2 - Structural Insight: Traditional NLP

Whilst word frequencies can help to explore the content of texts and be used in document classification, the order and wider context of words imparts additional meaning, which can be key to understanding research-based texts. However, enabling computers to understand human language, with all of its variations and irregular, evolving forms and rules is a complicated challenge. Computational techniques for the analysis of human language, a discipline called Natural Language Processing (NLP), typically involves a series of algorithms, each using linguistic rules to perform a discrete task and identify patterns in text.

An NLP pipeline starts with a set of texts of interest - known as a corpus (singular), or corpora (plural). This raw text is often referred to as unstructured data, as opposed to data in a table or other format ready for analysis. Many pipelines start with some form of tokenization, which splits text into smaller units, referred to as tokens. This might be paragraphs into sentences, or phrases into individual words, punctuation, or words into sub-words. One specific case of tokenization is sentence segmentation which aims to isolate individual sentences. This can be as simple as splitting at every period, or can employ more complex algorithms [7,8]. In languages without compound words, such as English, word tokenization can be as simple as separating whenever there is a space [3], but becomes more difficult in languages such as German or Mandarin where spacing may not be present between individual words. Note that sentence segmentation and tokenization can be used as preliminary steps before bag-of-words approaches, but we include them here as tokenizers are often designed using a mix of linguistic and computational considerations [9]. The ultimate goal of tokenization is to capture units of text that have discrete meaning for the given task, and thus may include rules for different punctuation marks, or algorithms that collapse multiple words into a single token. For example, the Penn Treebank [10] tokenizer uses regular expressions to tokenize individual sentences and is therefore able to split common English contractions (e.g. “don’t” becomes “do” and “n’t”).

Once tokens are identified, we can tag each word based on its part of speech (POS) (noun, verb, adjective, adverb, etc.) by using part-of-speech taggers. POS tagging can be achieved through different approaches including rule-based, probabilistic, or hybrids of the two [11]. Having tagged parts of speech can make additional pre-processing easier by helping identify multiple forms of word sharing a base form. For example, if one wants to identify instances of infectious disease, their interest likely covers all variants (or surface forms) of the word “infect”: infects, infecting, infection, infectious, infectivity, etc., and treating these words as independent may skew frequency-based analyses. Stemming can mitigate this by removing end characters from words to collapse down to a given base form or stem (in the above example, “infectious” becomes “infect”). However, stemming can often confound superficially similar yet unrelated words (e.g. *universal, university*, *universe* would all be stemmed to *univers*). Lemmatization considers a language’s full vocabulary, as well as the surrounding context of a word and their POS to collapse a word down to its canonical form, or lemma. For example, “infectious” again becomes “infect” but “better” becomes “good”, a relationship not captured by simple stemming. Once stemmed or lemmatized, stop word removal can also be considered (see Paradigm 1). For simple models, the choices made in lemmatizing and stop word removal can have a large influence on model outcomes [12].

Dependency parsing is another processing step enhanced by tagged parts of speech. These methods analyse grammatical structure to tag relationships between different words in a sentence. This allows for many downstream tasks such as identifying the subjects of sentences, who is doing what action, or where and when events are occuring. Similarly, part of speech tags can help when words are spelled the same but have different meanings depending on context (e.g. words used as either a verb or a noun: “bark”, “breed”, “control”, “flower”), or sets of words that represent a concept or a formal group. Sets of one or more words that comprise a single concept or entity can often be identified through Named Entity Recognition (NER) (also known as entity identification, entity chunking, or entity extraction) which is the process of identifying objects or concepts of a certain type, such as people, locations, events, etc. For example, rather than the more common use as a verb, the word ‘swallow’ could be recognized as an ‘organism’ entity in the phrase "most species of swallow feed on insects" based on the context of use. Developing NER tools typically requires these entities to be labelled, from which a model can be developed to predict these entities.

In biomedical text analysis, specialised NER tools have been developed to identify genes, proteins [13], and other terms used in medical language [14]. In ecology and evolution, NER has largely focused on identifying taxonomic names [15–19], though there are many other potential uses [1]. Once entities are recognized, coreference resolution aims at identifying all the terms in a given text that refers to a particular entity. For example, NER identifies the occurrence of a named entity (e.g., “animal”), but does not recognize when two different entities refer to the same thing (e.g. “swallow” and “*Hirundinidae”*). Coreference resolution includes finding and resolving different types of references, such as words that refer back to the subject (anaphora/cataphora), coreferring noun phrases across sentences, etc.

Beyond NER, there are myriad possibilities for tailoring NLP pipelines to work with text from specific fields. Entity linking is a method of disambiguation which assigns a unique identity to entities in a text and allows researchers to cross-reference entities with external databases or knowledge systems across diverse disciplines [20–26]. Note that coreference resolution does not necessarily use an external reference, whereas entity linking does. Once entities have been correctly identified and resolved, one may want to extract semantic relationships between entities (e.g. predator-prey, gene-protein). Similar to entity linking, Relation extraction (RE) is a core task in the field of Information Extraction (IE). By leveraging dependency and entity recognition as above, researchers can transform unstructured information in texts into structured relational or tabular data [27], ready for further analysis. Although RE is a complex task that can be done through many approaches [28] it will be an essential task for automating the creation, expansion, and maintenance of biodiversity datasets as the scientific literature grows [1].

[Figure 1 here]

Paradigm 3 - Language Model￼s: Deep Learning in NLP

In general terms, a language model is a method that predicts the next word in a sequence, given the previous words. There are two main types of language models: n-gram models and artificial neural network based models. N-gram models predict words from a fixed window of previous words, with probabilities estimated by counting the frequencies of these n-grams in a corpus [27]. N-gram language models can be simple and efficient, but suffer from poor generalisation to words and word sequences not seen in the training data – by predicting next words using only the words that immediately precede, n-gram models ignore useful information from more distant words in a text [29]. Neural language models are more complex and computationally expensive, but offer many unique advantages due to their ability to learn vector representations from the text itself. The process of self-supervised learning generates supervisory signals from the data itself, for example by masking some part of the input data and having the model try to predict it. Self-supervised learning enables large unlabelled corpora to be used for training models, and is now commonly used to train large language models that are then fine-tuned to perform more specific tasks through additional training. In this way, neural language models can learn complex features from unstructured data and capture subtle patterns which may be challenging for traditional NLP models.

Instead of treating words simply as indices in a vocabulary, neural network models represent words as vectors of numbers (also called word embeddings, a form of vector representation). A word embedding is a real-valued vector representation of a word that encodes meaning and preserves semantic proximity among words. While bag-of-words approaches like the TF-IDF can be used as numeric representations of text, neural embeddings are learned by iterating over a corpus of text and learning the association between the words. By assuming that terms with similar words before and after are semantically similar, embedding models learn vector representations of the words and their average contexts. In this way, distances between words in a continuous vector (measured by cosine similarity for example), indicate how similar the words are expected to be in their linguistic meaning (i.e., how replaceable one word might be with another).

NLP has seen a revolution starting with the incorporation of neural network based models in the 2010s, and with the development of new ways to represent text [30] that work well with deep learning methods [31,32]. Deep learning is now ubiquitous in NLP research, often superseding traditional NLP pipelines [33]. Various deep neural network (DNN) architectures have been used to address NLP tasks including convolutional neural networks (CNN), recurrent neural networks — most famously represented by the Long-Short Term Memory (LSTM) and the Bi-directional LSTM models — and graph neural networks [34]. Similar to first generation neural language models (word2vec, GloVe), these DNN-based models aim to produce meaningful word embeddings. The deep learning model that sparked the "NLP revolution" is the Bidirectional Encoder Representations from Transformers (BERT). BERT's main novelty is the use of a Transformer [31] as a neural architecture for encoding context in word embeddings [32]. Transformers are a family of encoder-decoder neural networks that use the mechanism of self-attention to capture long-range dependencies between distant elements in a sequence. In the context of NLP, self-attention enables Transformers to learn the context of a word relative to the entire input sequence. This contrasts with earlier approaches, including static embeddings (e.g. word2vec [30]), which represent a word as the same vector regardless of its context, and directional models (e.g. LSTM [35]), which only capture sequential dependencies (left-to-right and/or right-to-left). For example, word2vec will have the same embedding for the *bark* of a tree and the *bark* of a dog, while a Transformer will look at the surrounding words to generate an embedding that captures the meaning of *bark* based on the context in which it occurs. Unlike word2vec, which learns a single embedding per word, BERT breaks down words into subwords, allowing it to model embeddings for individual characters all the way through to complete words [9]. By combining subword embeddings, a BERT model can generate vector representations of out-of-vocabulary words, increasing its utility across different corpora and domains.

Following BERT's breakthrough, many alternative neural language models (e.g. XLNet, RoBERTa, ELECTRA) have been developed that use different architectures, hyperparameter tuning, pre-training objectives, and additional training data [36–38]. Since the pretraining of these language models does not require labelled data, there has been a trend of training increasingly large models using more and more massive unlabelled corpora. These large language models (LLMs) push computational boundaries through a mix of new neural architectures and sheer increases in size (number of parameters, size of vocabulary, number of learned vectors, all reflecting a large amount of pre-training literature) [33,39,40]. LLMs pre-trained on vast amounts of text often serve as foundation models that can be adapted to a wide range of downstream tasks through additional supervised learning with smaller datasets (Zhou et al. 2023) (see section *Transfer learning: adapting language models for ecology & evolution*). Foundation models are typically very large, with billions of parameters, and the past few years have seen rapid growth in the size of LLMs [33]. Such massive language models are computationally and financially expensive to train [39,41]. Several lines of work aim to create smaller language models with lower computational requirements but comparable performance [42–44] (e.g. ALBERT, DistilBERT, MobileBERT, TinyBERT, LLaMA, Mixtral). As the environmental impact of training and using large language models is not negligible, this should be considered when developing new NLP models [45].

Most large language models are pre-trained on general-domain corpora. For instance, the first iteration of BERT was trained on English Wikipedia (2,500M words) and the BookCorpus (800M words) [46]. Although general-domain LLMs pre-training corpora are derived from multiple massive sources, they may not be diverse enough to generalise to a specific domain or task that involves a specialised vocabulary, e.g. ecology and evolutionary biology. Pretraining to a specific domain can lead to increased performance in both high and low resource settings [47]. Some models have been pre-trained using scientific corpora such as SciBERT, trained on a random sample of 1.4 million papers from Semantic Scholar (Ammar et al. 2018; Beltagy, Lo, and Cohan 2019). BioBERT (Lee et al. 2020), a biomedical language model, expands on BERT using additional training data from PubMed abstracts and full-text articles, and PubMedBERT is trained on 14 million abstracts from PubMed [51]. These models have become the launch points for custom NLP pipelines for scientific and biomedical text (Neumann et al. 2019; Parmar et al. 2020), allowing for advances in building literature-based biomedical databases. Despite their demonstrated use in biomedical sciences, large language models are just beginning to be adopted in ecology and evolution [1,15,53,54], and to our knowledge there is currently only one large language model, BiodivBERT, trained explicitly on biodiversity-related texts [53].

Using language models in ecology & evolution

Text mining is a broad domain with an array of approaches that can be used for the same tasks. Above we outlined three broad paradigms (Frequency-based, Traditional NLP, and Deep learning). While deep learning-based language models are revolutionary and an extremely active area of research, methods and tools from each of the three paradigms can be effective means to analyse text in ecology and evolution, even if not considered the current “state of the art”. Often simpler rule-based tools can be more computationally efficient, and perform better for a specific task than a more general, probabilistic machine-learning based approach. Further, three paradigms can interface with each other. For example, rule-based methods can be used for generating labelled data for supervised training of machine learning models. More generally, the output of one type of task can be used as inputs to the next (Fig. 2). For example Kulkarni and Di Minin (2021) and Hunter et al. (2023) use TF-IDF to find and remove duplicate articles, then use a hashing function to turn texts into features for use in a neural network based classification model. However, as large language models are likely to provide the greatest flexibility and potential for future development, we focus the rest of the article on their application, highlighting challenges and solutions to developing ecology-specific language models, gathering suitable training data, and identifying avenues for prompt-based interaction with LLMs.

[Figure 2 here]

State-of-the-art in ecology & evolution: learning in a low-resource domain

Despite many promising applications of NLP in ecology [1,18], we currently have few domain-specific NLP tools [15,53]. This is in part because ecology is a low-resource domain [57], meaning there are few large open-access text databases available for training foundation models [1], and there are few gold standard databases with task-specific labels needed for supervised learning. Some ecology-specific gold standard datasets exist, such as COPIOUS, a corpus directed towards the extraction of species occurrences from the biodiversity literature [58], and the newer BiodivNERE [59], a gold standard corpus for NER and RE. While gold standard datasets are essential for comparing the performance of new approaches as they are developed, they do not currently provide enough examples to train deep learning NLP models from scratch. A major barrier for text mining in ecology and evolution is the lack of data formatted for the training, testing, and development of domain-specific NLP tools. Even with a concerted effort, it is unlikely that we will reach the same quantities of annotated data that are available for other deep learning-based ecological tasks such as computer vision to identify species in camera traps [60,61].

Gathering labelled training data for language models typically involves manual tagging of documents, phrases, or words by expert annotators. Given the increasing availability of computing power, expert labelling and curation of training data can be the most expensive and time-consuming component of a custom NLP model. The cost is often high because of the manual effort of (potentially highly trained) humans. When tasks are difficult or require domain expertise, there may be trade-offs between the annotators' time and number of examples needed for training, the need for multiple annotators to limit individual biases and provide a consensus annotation, or crowdsourcing annotations to produce training sets at the risk of poorer quality or less reliability among annotators. As an alternative to manual curation, we can use a mixture of simple rules based on the content of existing databases to increase the efficiency of labelling [62]. In ecology & evolution, comparative databases of species traits, biodiversity responses, and experimental data built from primary literature [63,64] offer excellent resources for creating labelled training data. These are typically manually constructed by multiple contributors, have undergone some quality checks, and represent decades of effort curating facts and observations from literature.

When biodiversity databases can be paired with source texts, approaches such as data augmentation and weak supervision can be used to efficiently generate labelled data [57]. Data augmentation involves artificially expanding a dataset by applying transformations to existing labelled data. Widely used in computer vision, transformations which preserve the semantics of the images, such as image rotation, flipping or cropping, are applied to a dataset to construct additional synthetic data. This technique has only recently been adopted in NLP because the nature of text makes it more difficult to derive transformations that maintain the semantics of the original sentences, however new techniques for data augmentation in NLP are being developed [65,66]. Unlike data augmentation, which requires an initial set of labelled data, weak supervision uses unlabelled texts for which labels are obtained (semi-)automatically from simpler rule-based methods (e.g. outputs of paradigms 1 and 2) or pattern matching based on external information sources [62]. A popular form of weak supervision is distant supervision, in which labels are derived from existing knowledge bases, dictionaries, and other forms of structured knowledge sources [67]. The main challenge with weak supervision is reducing the noise and inaccuracies introduced by these imperfect annotation sources.

Transfer learning: adapting existing language models

If only a small set of labelled data are available, it may not be possible to train a deep learning model from scratch. An alternative approach is to fine-tune existing large language models (trained either on general or scientific texts) to a target task via transfer learning. This process reduces the need for labelled data as it uses fewer examples than would be needed to train a full language model. Through transfer learning a variety of NLP tasks could feasibly be performed by a single neural language model. Thus, LLMs allow flexible development of custom pipelines without specialised linguistic knowledge required to build traditional NLP tools. While there exist pre-trained language models designed to work with scientific literature, they are either explicitly biomedical focused, such as bioBERT [50] and PubMedBERT [51], or the majority of the training corpus is from the biomedical domain, e.g. scispaCy [13] and sciBERT [49]. The development of language models specifically for ecology and evolution, or disciplines that bridge ecology and biomedical sciences, such as veterinary medicine, could foreseeably be achieved by biomedical or more generalised language models as a foundation. A recent example is TaxoNERD [15], a model for taxonomic named entity recognition trained with transfer learning and updating of the bioBERT biomedical language model. While transfer learning offers a viable avenue for adapting LLMs for tasks in ecology and evolution, it requires the curation of labelled training data specific for each task.

Generative LLMs and prompt-based interaction for complex tasks

The training and fine-tuning of LLMs is a computationally expensive task and relies on the availability of large amounts of training data. Developing domain specific language models may be necessary for specialised tasks, but as language models are adapted to answer questions and respond directly to Natural Language prompts, it may be more efficient to directly prompt LLMs to identify, extract, and harmonise data directly from scientific literature [54]. As an alternative to training domain-specific models, prompt-based learning with a pre-trained, general-purpose LLM allows for the execution of domain-specific tasks without the need for pre-training or transfer learning [68]. By additionally training LLMs to act as generative question answering models (e.g. OpenAI's ChatGPT built on GPT, and Meta's Llama-2) [69,70], researchers can directly prompt models to perform more complex information extraction tasks [71], as been demonstrated for identifying potential pest controllers [54], and the locations of species observations [72]. In this context a ’prompt’ refers to a piece of text to instruct a language model to perform a task in natural language rather than through computer code. Prompts may include explicit examples of the task to be completed (few-shot learning) or be composed of an instruction only (zero-shot learning). Prompt-based learning has become a popular area of research in recent years [73], with applications across multiple scientific disciplines, including text classification [73–75], NER [75,76] and relation extraction [74,77–81]. While obtaining reliable performance of LLMs on complex tasks is challenging and may require extensive experimentation, there is a rapidly growing repertoire of prompting techniques that show promise.

Least-to-most prompting [82] proposes to decompose a complex task into a series of subtasks, each as their own prompt, which are then queried in sequence. For example, in a task that involves both NER and RE, one could first prompt the LLM with an NER task (e.g., “list all species mentioned in the text”), and then detail the RE task and any other instructions (e.g., “return a table with two columns: If a listed species acts as a parasite of another species mentioned in the text [based on criteria x, y, z], then return this parasite in column one and its host(s) in column two”) [54]. Least-to-most prompting has been shown to improve the ability of GPT-3 to solve difficult reasoning problems [82]. Similarly, chain-of-thought prompting uses a series of examples in the prompt to show intermediate reasoning, which aids the model in decision making and can improve performance on a variety of reasoning tasks [78]. These techniques may be improved by allowing the model to sample multiple reasoning paths and generate a result based on the most consistent answer [79,80]. Such approaches are likely to be helpful in ecology to simultaneously extract entities and make decisions about their ecological roles (e.g., as pest or parasite) and relationships to other entities (e.g., host-parasite or prey-predator relations) [54]. While prompt-based interaction can be quite useful, the construction of an effective prompt may be highly task- and data-dependent, and thus may require extensive fine-tuning to account for intricacies, ambiguities, or exceptions. Some of the difficulties of effective prompt engineering may be circumvented by fine-tuning the LLM parameters through transfer learning, however, prompt based approaches may offer a substantial reduction in computational cost if using existing LLMs.

Conclusion

Natural Language Processing is a rapidly developing field which offers a diversity of approaches useful to researchers in ecology and evolution. Deep learning has revolutionised NLP approaches and now allows researchers without extensive linguistics or computational training to employ them. In particular, the proliferation of generative LLMs will likely foster a wide adoption of text mining in ecology, especially for complex tasks related to data extraction and synthesis of biodiversity data. However, given the energy and computational resources needed to train and implement deep-learning based language models, we emphasise that many of the simpler rule-based and traditional NLP approaches reviewed here may have comparable performance while requiring fewer resources. Currently in ecology and evolution our ability to develop state-of-the-art text mining tools is limited by the lack of open access to full texts of scientific publications [1]. Relevant to this is that university libraries are fundamentally changing – we now have "digital librarians" and ecologists can advocate for this type of science to be supported through their libraries and contracts with publishers.

We are also in the midst of an exponential growth in size and number of language models, both open and closed-source. With closed-source models such as OpenAI's ChatGPT using subscription-based payments for access to newest models and pay-per-token for legacy models which may become unavailable with limited notice, the use of closed-source models or software reliant on them will create an unprecedented challenge for reproducible research [83]. Research code is already suffering from issues of reproducibility [84], and the adoption of closed-source models which lack transparency in their training data and/or model architecture directly opposes the open science movement. Competitive open source LLMs are being released, but these do not mitigate the resource and environmental costs of LLMs (e.g. at least 400GB of memory is needed to swiftly run Falcon-180 billion parameter model [85]).

Prompt-based interaction with generative LLMs can increase the efficiency of data synthesis, but there is a risk that these models generate false or harmful information [39]. However, ecology and evolution have a rich history of literature-based data synthesis which can be leveraged for fine-tuning foundation models and validating outputs. Further, the potential for LLMs to interact with external databases through Retrieval Augmented Generation shows great promise for reducing false information [86] and building models that can keep up with rapidly expanding literature. Significant open questions for the future remain - How important is it to develop field-specific language models versus using general LLMs or transfer learning from foundation models from related disciplines? How can we ensure the tools we adopt will be open access and developed using transparent methods? Should we attempt to limit environmental impacts of training competing large-scale models, instead using simpler approaches or transfer learning where possible? Regardless of how this methodological landscape develops, it will be important to keep the “human in the loop”, from sense-checking models and training data, identifying harmful or inaccurate outputs of generative models, balancing model complexity with environmental impacts, and ensuring the reproducible, open, and equitable access of literature, data, code, and models.

Figures

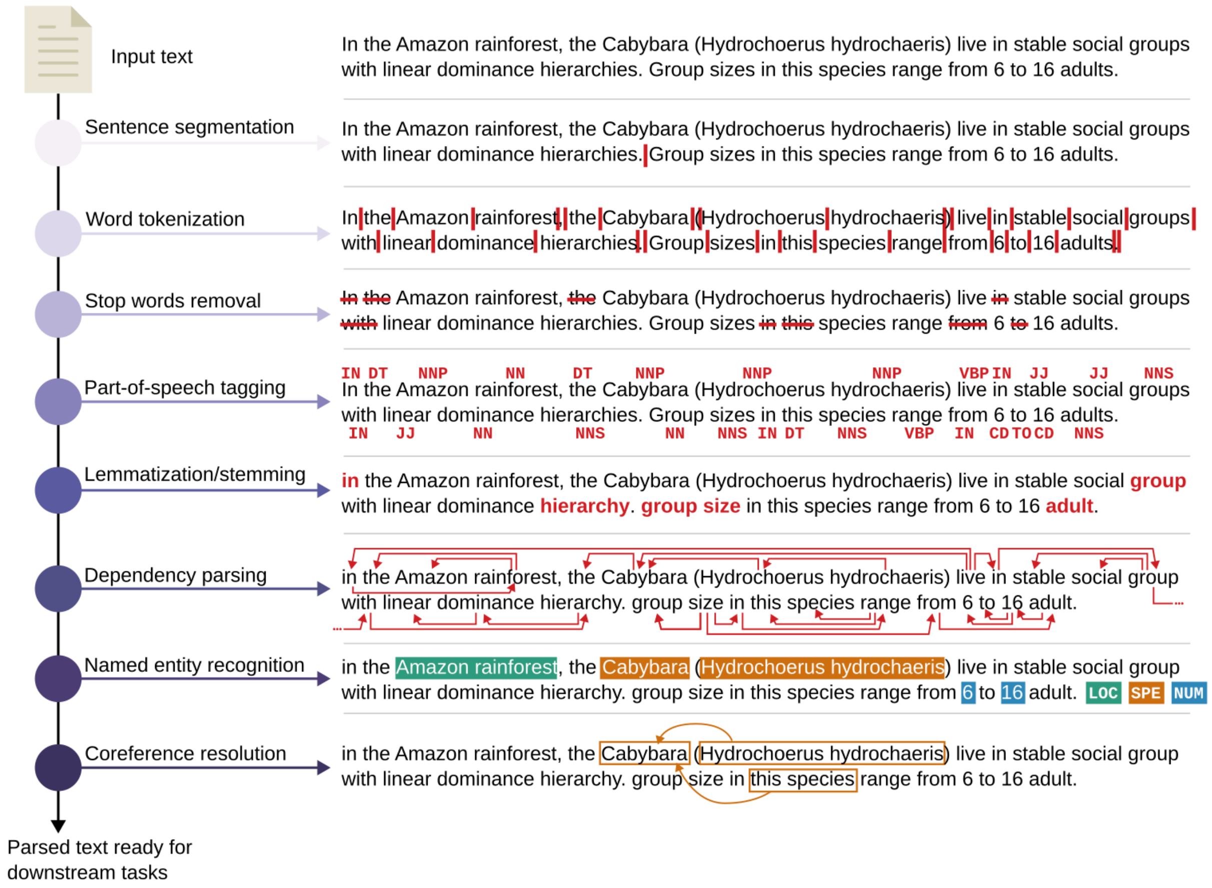
Input text
In the Amazon rainforest, the Cabybara (Hydrochoerus hydrochaeris) live in stable social groups with linear dominance hierarchies. Group sizes in this species range from 6 to 16 adults.
Sentence segmentation
Word tokenization
Stop words removal
Part-of-speech tagging
IN DT NNP NN DT NNP NNP NNP VBP IN JJ JJ NNS
IN JJ NN NNS NN NNS IN DT NNS VBP IN CD TO CD NNS
Lemmatization/stemming
in the Amazon rainforest, the Cabybara (Hydrochoerus hydrochaeris) live in stable social group with linear dominance hierarchy. group size in this species range from 6 to 16 adult.
Dependency parsing
Named entity recognition
LOC
SPE
NUM
Coreference resolution
Parsed text ready for downstream tasks

Figure 1: Illustration of common text processing tasks, including sentence segmentation, stop words removal, tokenization, POS-tagging, lemmatization / stemming, dependency parsing, Named Entity Recognition, and coreference resolution. POS-tagging involves marking up tokens with a set of descriptive POS tags, e.g. determiner (DT), proper noun (NNP), adjective (JJ)... Dependency parsing creates a tree-like representation of the grammatical relationships between words in a sentence. Note that the order and inclusion of individual steps in a pipeline will depend on the task. Example text derived from Herrera et al. (2011).

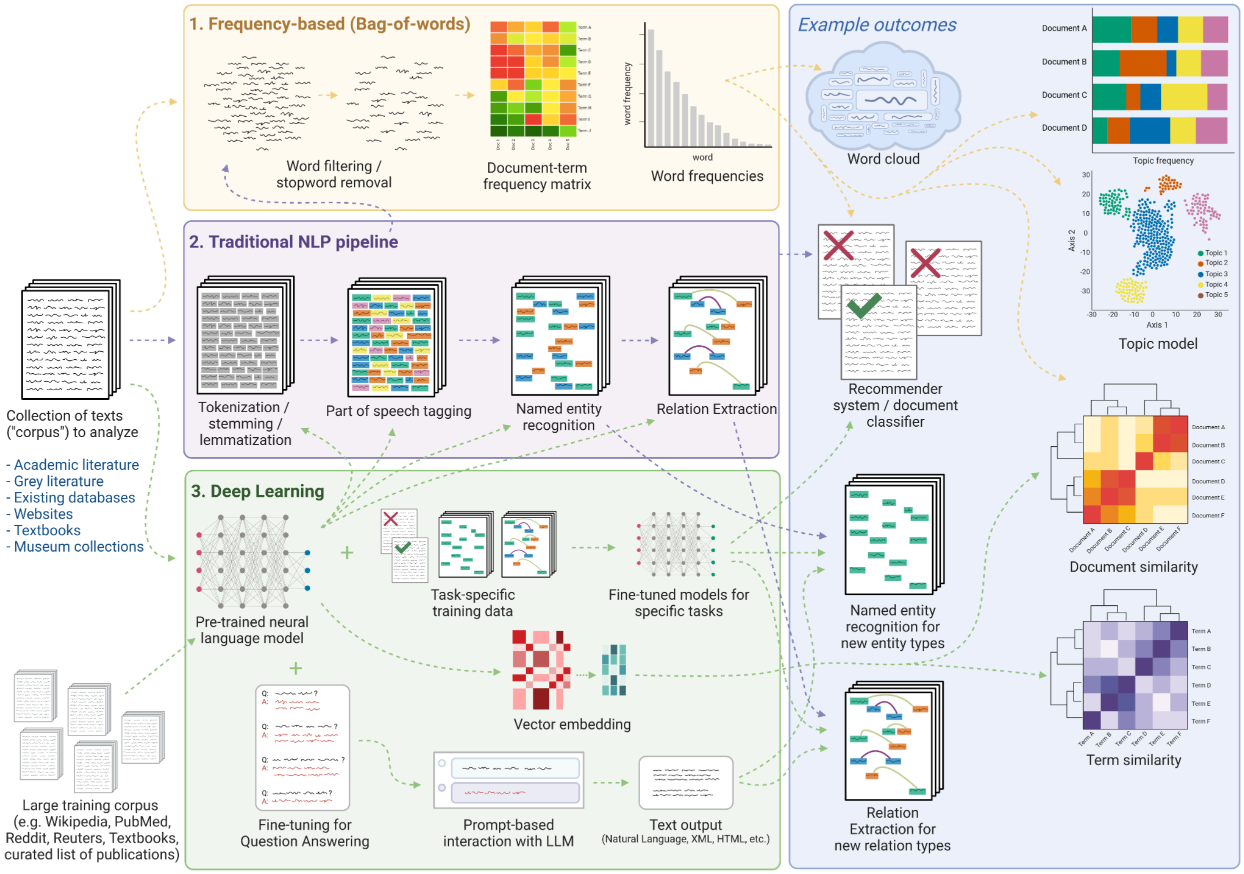

Figure 2: Example steps for three text mining paradigms: 1. Frequency-based (bag-of-words) (frequency-based) approach, 2. Traditional NLP pipeline, and 3. Deep learning based language models. Dashed arrows indicate possible interactions between each of the paradigms (for example, text pre-processed using a classical NLP pipeline could be analysed using bag-of-words approaches, or fed into a deep-learning based document classifier). The right panel provides some examples of outcomes, including quantifying document similarity, topic modelling, and training models for document classification, Named Entity Recognition, and Relation Extraction. Note that for steps in the Traditional NLP section there are often additional task-specific external data sources (e.g. word lists, dictionaries, labelled training data) not depicted here.

References

1. Farrell MJ, Brierley L, Willoughby A, Yates A, Mideo N. 2022 Past and future uses of text mining in ecology and evolution. *Proc. R. Soc. B Biol. Sci.* **289**, 20212721. (doi:10.1098/rspb.2021.2721)
2. Brandies PA, Hogg CJ. 2021 Ten simple rules for getting started with command-line bioinformatics. *PLOS Comput. Biol.* **17**, e1008645. (doi:10.1371/journal.pcbi.1008645)
3. Bird S, Klein E, Loper E. 2009 *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.
4. Honnibal M, Montani I. 2017 spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.
5. Nunez‐Mir GC, Iannone BV, Pijanowski BC, Kong N, Fei S. 2016 Automated content analysis: addressing the big literature challenge in ecology and evolution. *Methods Ecol. Evol.* **7**, 1262–1272. (doi:10.1111/2041-210X.12602)
6. Stringham OC, Moncayo S, Hill KGW, Toomes A, Mitchell L, Ross JV, Cassey P. 2021 Text classification to streamline online wildlife trade analyses. *PLOS ONE* **16**, e0254007. (doi:10.1371/journal.pone.0254007)
7. Kiss T, Strunk J. 2006 Unsupervised Multilingual Sentence Boundary Detection. *Comput. Linguist.* **32**, 485–525. (doi:10.1162/coli.2006.32.4.485)
8. Honnibal M, Johnson M. 2015 An Improved Non-monotonic Transition System for Dependency Parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1373–1378. Lisbon, Portugal: Association for Computational Linguistics. (doi:10.18653/v1/D15-1162)
9. Mielke SJ *et al.* 2021 Between words and characters: A Brief History of Open-Vocabulary Modeling and Tokenization in NLP. *arXiv:2112.10508*
10. Marcus MP, Marcinkiewicz MA, Santorini B. 1993 Building a large annotated corpus of English: the penn treebank. *Comput. Linguist.* **19**, 313–330.
11. Kanakaraddi SG, Nandyal SS. 2018 Survey on Parts of Speech Tagger Techniques. In *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)*, pp. 1–6. (doi:10.1109/ICCTCT.2018.8550884)
12. HaCohen-Kerner Y, Miller D, Yigal Y. 2020 The influence of preprocessing on text classification using a bag-of-words representation. *PLOS ONE* **15**, 1–22. (doi:10.1371/journal.pone.0232525)
13. Neumann M, King D, Beltagy I, Ammar W. 2019 ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. *Proc. 18th BioNLP Workshop Shar. Task* , 319–327. (doi:10.18653/v1/W19-5034)
14. Perera N, Dehmer M, Emmert-Streib F. 2020 Named Entity Recognition and Relation Detection for Biomedical Information Extraction. *Front. Cell Dev. Biol.* **8**.
15. Le Guillarme N, Thuiller W. 2022 TaxoNERD: Deep neural models for the recognition of taxonomic entities in the ecological and evolutionary literature. *Methods Ecol. Evol.* **13**, 625–641. (doi:10.1111/2041-210X.13778)
16. Gerner M, Nenadic G, Bergman CM. 2010 LINNAEUS: A species name identification system for biomedical literature. *BMC Bioinformatics* **11**, 85. (doi:10.1186/1471-2105-11-85)
17. Thessen AE, Parr CS. 2014 Knowledge Extraction and Semantic Annotation of Text from the Encyclopedia of Life. *PLoS ONE* **9**, e89550. (doi:10.1371/journal.pone.0089550)
18. Thessen AE, Cui H, Mozzherin D. 2012 Applications of Natural Language Processing in Biodiversity Science. *Adv. Bioinforma.* **2012**, 391574. (doi:10.1155/2012/391574)
19. Mozzherin D, Myltsev A, Zalavadiya H. 2023 gnames/gnfinder: v1.1.3. (doi:10.5281/zenodo.8104625)

20. Hassell J, Aleman-Meza B, Arpinar IB. 2006 Ontology-Driven Automatic Entity Disambiguation in Unstructured Text. In *The Semantic Web - ISWC 2006* (eds I Cruz, S Decker, D Allemang, C Preist, D Schwabe, P Mika, M Uschold, LM Aroyo), pp. 44–57. Berlin, Heidelberg: Springer. (doi:10.1007/11926078_4)
21. Volz R, Kleb J, Mueller W. 2007 Towards Ontology-based Disambiguation of Geographical Identifiers. I3.
22. Bodenreider O. 2008 Biomedical ontologies in action: role in knowledge management, data integration and decision support. *Yearb. Med. Inform.* , 67–79.
23. Alexopoulou D, Andreopoulos B, Dietze H, Doms A, Gandon F, Hakenberg J, Khelif K, Schroeder M, Wächter T. 2009 Biomedical word sense disambiguation with ontologies and metadata: automation meets accuracy. *BMC Bioinformatics* **10**, 28. (doi:10.1186/1471-2105-10-28)
24. Mungall CJ, Torniai C, Gkoutos GV, Lewis SE, Haendel MA. 2012 Uberon, an integrative multi-species anatomy ontology. *Genome Biol.* **13**, R5. (doi:10.1186/gb-2012-13-1-r5)
25. Dececchi TA, Balhoff JP, Lapp H, Mabee PM. 2015 Toward Synthesizing Our Knowledge of Morphology: Using Ontologies and Machine Reasoning to Extract Presence/Absence Evolutionary Phenotypes across Studies. *Syst. Biol.* **64**, 936–952. (doi:10.1093/sysbio/syv031)
26. Le Guillarme N *et al.* 2023 The Soil Food Web Ontology: Aligning trophic groups, processes, resources, and dietary traits to support food-web research. *Ecol. Inform.* **78**, 102360. (doi:10.1016/j.ecoinf.2023.102360)
27. Jurafsky D, Martin JH. 2023 *Speech and Language Processing*. 3rd edn. Pearson Education.
28. Detroja K, Bhensdadia CK, Bhatt BS. 2023 A survey on Relation Extraction. *Intell. Syst. Appl.* **19**, 200244. (doi:10.1016/j.iswa.2023.200244)
29. Bengio Y, Ducharme R, Vincent P, Jauvin C. 2003 A Neural Probabilistic Language Model. *J. Mach. Learn. Res.* **3**, 1137–1155.
30. Mikolov T, Chen K, Corrado G, Dean J. 2013 Efficient Estimation of Word Representations in Vector Space. *ArXiv13013781 Cs*
31. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. 2017 Attention Is All You Need. *ArXiv170603762 Cs*
32. Devlin J, Chang M-W, Lee K, Toutanova K. 2019 BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805*
33. Zhao WX *et al.* 2023 A Survey of Large Language Models.
34. Wu L, Chen Y, Shen K, Guo X, Gao H, Li S, Pei J, Long B. 2021 Graph Neural Networks for Natural Language Processing: A Survey. *arXiv:2106.06090*
35. Hochreiter S, Schmidhuber J. 1997 Long Short-Term Memory. *Neural Comput.* **9**, 1735–1780. (doi:10.1162/neco.1997.9.8.1735)
36. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov R, Le QV. 2020 XLNet: Generalized Autoregressive Pretraining for Language Understanding.
37. Liu Y *et al.* 2019 RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692*
38. Clark K, Luong M-T, Le QV, Manning CD. 2020 ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators.
39. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. 2021 On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 610–623. Virtual Event Canada: ACM. (doi:10.1145/3442188.3445922)
40. Zhou C *et al.* 2023 A Comprehensive Survey on Pretrained Foundation Models: A History

from BERT to ChatGPT.

41. Luccioni AS, Viguier S, Ligozat A-L. 2022 Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model. *arXiv:2211.02001*
42. Sanh V, Debut L, Chaumond J, Wolf T. 2020 DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv:1910.01108*
43. Jiang AQ *et al.* 2024 Mixtral of Experts. *arXiv:2401.04088* (doi:10.48550/arXiv.2401.04088)
44. Touvron H *et al.* 2023 LLaMA: Open and Efficient Foundation Language Models. *arXiv:2302.13971* (doi:10.48550/arXiv.2302.13971)
45. Everman B, Villwock T, Chen D, Soto N, Zhang O, Zong Z. 2023 Evaluating the carbon impact of large language models at the inference stage. In *2023 IEEE international performance, computing, and communications conference (IPCCC)*, pp. 150–157. (doi:10.1109/IPCCC59175.2023.10253886)
46. Zhu Y, Kiros R, Zemel R, Salakhutdinov R, Urtasun R, Torralba A, Fidler S. 2015 Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books. *arXiv:1506.06724* (doi:10.48550/ARXIV.1506.06724)
47. Gururangan S, Marasović A, Swayamdipta S, Lo K, Beltagy I, Downey D, Smith NA. 2020 Don’t stop pretraining: Adapt language models to domains and tasks. *arXiv:2004.10964*
48. Ammar W *et al.* 2018 Construction of the Literature Graph in Semantic Scholar. *arXiv:1805.02262*
49. Beltagy I, Lo K, Cohan A. 2019 *SciBERT: A Pretrained Language Model for Scientific Text*.
50. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. 2020 BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**, 1234–1240.
51. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, Naumann T, Gao J, Poon H. 2022 Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Trans. Comput. Healthc.* **3**, 1–23. (doi:10.1145/3458754)
52. Parmar J, Koehler W, Bringmann M, Volz KS, Kapicioglu B. 2020 Biomedical Information Extraction for Disease Gene Prioritization. *arXiv:2011.05188*
53. Abdelmageed N, Löffler F, König-Ries B. 2023 BiodivBERT: a pre-trained language model for the biodiversity domain. pp. 62–71. CEUR-WS.org.
54. Scheepens D, Millard J, Farrell M, Newbold T. 2024 Large language models help facilitate the automated synthesis of information on potential pest controllers. *bioRxiv* , 2024.01.12.575330. (doi:10.1101/2024.01.12.575330)
55. Kulkarni R, Di Minin E. 2021 Automated retrieval of information on threatened species from online sources using machine learning. *Methods Ecol. Evol.* **12**, 1226–1239. (doi:10.1111/2041-210X.13608)
56. Hunter SB, Mathews F, Weeds J. 2023 Using hierarchical text classification to investigate the utility of machine learning in automating online analyses of wildlife exploitation. *Ecol. Inform.* **75**, 102076. (doi:10.1016/j.ecoinf.2023.102076)
57. Hedderich MA, Lange L, Adel H, Strötgen J, Klakow D. 2021 A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios. pp. 2545–2568. Online: Association for Computational Linguistics. (doi:10.18653/v1/2021.naacl-main.201)
58. Nguyen N, Gabud R, Ananiadou S. 2019 COPIOUS: A gold standard corpus of named entities towards extracting species occurrence from biodiversity literature. *Biodivers. Data J.* **7**, e29626. (doi:10.3897/BDJ.7.e29626)
59. Abdelmageed N, Löffler F, Feddoul L, Algergawy A, Samuel S, Gaikwad J, Kazem A, König-Ries B. 2022 BiodivNERE: Gold standard corpora for named entity recognition and relation extraction in the biodiversity domain. *Biodivers. Data J.* **10**, e89481. (doi:10.3897/BDJ.10.e89481)

60. Swanson A, Kosmala M, Lintott C, Simpson R, Smith A, Packer C. 2015 Snapshot Serengeti, high-frequency annotated camera trap images of 40 mammalian species in an African savanna. *Sci. Data* **2**, 150026. (doi:10.1038/sdata.2015.26)
61. Norouzzadeh MS, Nguyen A, Kosmala M, Swanson A, Palmer MS, Packer C, Clune J. 2018 Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proc. Natl. Acad. Sci.* **115**, E5716–E5725. (doi:10.1073/pnas.1719367115)
62. Ratner A, Bach SH, Ehrenberg H, Fries J, Wu S, Ré C. 2020 Snorkel: rapid training data creation with weak supervision. *VLDB J.* **29**, 709–730. (doi:10.1007/s00778-019-00552-1)
63. Heberling JM, Miller JT, Noesgaard D, Weingart SB, Schigel D. 2021 Data integration enables global biodiversity synthesis. *Proc. Natl. Acad. Sci.* **118**. (doi:10.1073/pnas.2018093118)
64. Poisot T, Bruneau A, Gonzalez A, Gravel D, Peres-Neto P. 2019 Ecological Data Should Not Be So Hard to Find and Reuse. *Trends Ecol. Evol.* **34**, 494–496. (doi:10.1016/j.tree.2019.04.005)
65. Feng SY, Gangal V, Wei J, Chandar S, Vosoughi S, Mitamura T, Hovy E. 2021 A Survey of Data Augmentation Approaches for NLP. *arXiv:2105.03075*
66. Chen J, Tam D, Raffel C, Bansal M, Yang D. 2023 An Empirical Survey of Data Augmentation for Limited Data Learning in NLP. *Trans. Assoc. Comput. Linguist.* **11**, 191–211. (doi:10.1162/tacl_a_00542)
67. Mintz M, Bills S, Snow R, Jurafsky D. 2009 Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, pp. 1003–1011. USA: Association for Computational Linguistics.
68. Gao T, Fisch A, Chen D. 2021 Making Pre-trained Language Models Better Few-shot Learners. *arXiv:2012.15723* (doi:10.48550/arXiv.2012.15723)
69. Ouyang L *et al.* 2022 Training language models to follow instructions with human feedback. *arXiv:2203.02155*
70. Touvron H *et al.* 2023 Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv:2307.09288*
71. Wei X *et al.* 2023 Zero-Shot Information Extraction via Chatting with ChatGPT. *arXiv:2302.10205*
72. Castro A, Pinto J, Reino L, Pipek P, Capinha C. 2024 Large language models overcome the challenges of unstructured text data in ecology. *bioRxiv* , 2024.01.23.576654. (doi:10.1101/2024.01.23.576654)
73. Mayer CWF, Ludwig S, Brandt S. 2023 Prompt text classifications with transformer models! An exemplary introduction to prompt-based learning with large language models. *J. Res. Technol. Educ.* **55**, 125–141. (doi:10.1080/15391523.2022.2142872)
74. Chen Q *et al.* 2023 An extensive benchmark study on biomedical text generation and mining with ChatGPT. *Bioinformatics* **39**, btad557. (doi:10.1093/bioinformatics/btad557)
75. Wu S, Irsoy O, Lu S, Dabravolski V, Dredze M, Gehrmann S, Kambadur P, Rosenberg D, Mann G. 2023 BloombergGPT: A Large Language Model for Finance. *arXiv:2303.17564* (doi:10.48550/arXiv.2303.17564)
76. Hu Y *et al.* 2024 Improving Large Language Models for Clinical Named Entity Recognition via Prompt Engineering. *arXiv:2303.16416* (doi:10.48550/arXiv.2303.16416)
77. He K, Huang Y, Mao R, Gong T, Li C, Cambria E. 2023 Virtual prompt pre-training for prototype-based few-shot relation extraction. *Expert Syst. Appl.* **213**, 118927. (doi:10.1016/j.eswa.2022.118927)
78. Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, Chi E, Le Q, Zhou D. 2023 Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *arXiv:2201.11903*

(doi:10.48550/arXiv.2201.11903)

79. Wang X, Wei J, Schuurmans D, Le Q, Chi E, Narang S, Chowdhery A, Zhou D. 2023 Self-Consistency Improves Chain of Thought Reasoning in Language Models. *arXiv:2203.11171* (doi:10.48550/arXiv.2203.11171)
80. Li Y, Lin Z, Zhang S, Fu Q, Chen B, Lou J-G, Chen W. 2023 Making Large Language Models Better Reasoners with Step-Aware Verifier. *arXiv:2206.02336* (doi:10.48550/arXiv.2206.02336)
81. Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y. 2023 Large Language Models are Zero-Shot Reasoners. *arXiv:2205.11916* (doi:10.48550/arXiv.2205.11916)
82. Zhou D *et al.* 2023 Least-to-Most Prompting Enables Complex Reasoning in Large Language Models. *arXiv.2205.10625* (doi:10.48550/arXiv.2205.10625)
83. Jenkins GB *et al.* 2023 Reproducibility in ecology and evolution: Minimum standards for data and code. *Ecol. Evol.* **13**, e9961. (doi:10.1002/ece3.9961)
84. Trisovic A, Lau MK, Pasquier T, Crosas M. 2022 A large-scale study on research code quality and execution. *Sci. Data* **9**, 60. (doi:10.1038/s41597-022-01143-6)
85. 2023 tiiuae/falcon-180B · Hugging Face. See https://huggingface.co/tiiuae/falcon-180B (accessed on 13 November 2023).
86. Lewis P *et al.* 2021 Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. (doi:10.48550/arXiv.2005.11401)
87. Herrera EA, Salas V, Congdon ER, Corriale MJ, Tang-Martínez Z. 2011 Capybara social structure and dispersal patterns: variations on a theme. *J. Mammal.* **92**, 12–20. (doi:10.1644/09-MAMM-S-420.1)
88. Schneider FD *et al.* 2019 Towards an ecological trait-data standard. *Methods Ecol. Evol.* **10**, 2006–2019. (doi:10.1111/2041-210X.13288)

Table SM 1: Glossary of Key Terms and Concepts (alphabetical)

Accuracy: a common metric used to evaluate the performance of a machine learning model. Accuracy measures how often the model correctly predicts the outcomes within a given dataset. It is calculated by dividing the number of correct predictions by the total number predictions.

Anaphora/cataphora: a word or phrase used as a substitute for a word or phrase used earlier (anaphora) or later (cataphora) in a text. These are used to avoid repetition (e.g. it, they, their, some. For example, in "In the second week after hatching eagle chicks can hold up *their* heads", "their" is an anaphor, referring back to the eagle chicks.)

Artificial neural network: a type of machine learning model comprising a collection of connected units or nodes called artificial neurons organised into layers. Each connection between neurons has an associated weight. The process of training a neural network involves feeding it labelled training data, and adjusting the weights and biases during each iteration to minimise the difference between the predicted outputs and the true labels. This process is often done using optimization algorithms such as gradient descent.

"Bag of words" (BOW) approaches: methods that treat words (or n-grams) as independent entities, ignoring their common linguistic function (i.e. part of speech), or relative position in a sentence. Examples include word clouds, TF-IDF, and topic modelling.

Bidirectional Encoder Representations from Transformers (BERT): an open source machine learning framework developed by Google [32] based on the transformer deep learning architecture [31]. BERT models advanced previous neural language models by using self-attention mechanisms to understand the relationships between words in a sentence. BERT considers all words in a sentence at once, making it better at capturing context and meaning of given words. Also, by pre-training on large amounts of data rather than being trained from scratch, BERT became the first (?) general-purpose language model, allowing it to be fine-tuned for a wide variety of tasks.

Corpus: a collection of texts. Plural *corpora* refers to multiple collections of texts.

Data augmentation: a technique commonly used in machine learning to improve the generalisation and robustness of a machine learning model. The idea is to increase the diversity of a training dataset by generating several slightly-modified copies of existing data.

Deep learning: a subset of machine learning based on artificial neural networks. Inspired by the human brain, these models use multiple layers of processing to analyse progressively more abstract higher level features from data. This allows deep learning models to recognize and simulate complex patterns in images, text, sound, and other types of data.

Disambiguation: the process of determining which sense of a word is meant in a given context. For example, "bark" could refer to the bark of a tree, or the sound dogs make.

Distant supervision: a technique used in the field of machine learning, particularly in NLP and IE, which consists of using external knowledge bases to automatically label a large amount of data.

Encoder-decoder neural network: a type of deep neural network architecture that involves a two stage process. First input data are encoded into a numerical representation (typically a fixed length vector, often referred to as a hidden or latent representation). The decoder then takes this latent representation and produces an output in the desired format. This architecture is often used when the input is sequence data and the output is another type of sequence data. Different types of neural networks such as RNN, LSTM, CNN use this architecture.

Feature engineering: the process of using domain knowledge, curation, and data transformation to create predictor variables for predictive models.

Few-shot learning: a machine learning paradigm that focuses on training models with only a small number of examples for each class or task. The term “shot” refers to the number of examples provided for each class during training. Extreme scenarios of few-shot learning include one-shot learning and zero-shot learning.

Foundation model: a large, pre-trained language model that is not trained for specific tasks but rather aims to learn general language representations that can be fine-tuned for specific downstream tasks with relatively small amounts of task-specific labelled data.

Gold standard: a set of data that have been manually curated and verified to be the most accurate or correct for a given analysis. In NLP the gold standard typically represents a set of documents that have been expertly annotated and against which different models can be trained or tested.

Information retrieval (IR): the process of searching for and identifying documents that are relevant for a given task.

Language model: a type of statistical or machine learning model designed to capture the inherent structure, patterns, and semantics of one or several human languages. A language model is generally trained to predict which word is more likely to appear next in a sentence based on the previous words.

Large language model: a deep neural network-based language model trained on massive amounts of text using self-supervision. The largeness of these models is a key factor in their ability to generalise well across a wide range of language-related tasks. GPT-3, the model behind the popular chatbot ChatGPT, is an example of a large language model.

Latent Dirichlet Allocation (LDA): a generative statistical model for the unsupervised classification of documents. One of the most popular methods for topic modelling, LDA aims to find hidden topics that a document belongs to based on the words it contains.

Lemma & Lemmatization: a lemma is the canonical form of a word (e.g. the lemma of eats, eating, eaten, and ate is "eat"). Lemmatization is the process of determining the lemma of a word based on context and intended meaning.
N-grams: a set of N words that appear next to one another. For example, in the text "big oak tree", "big", "oak", and "tree" are unigrams, "big oak" and "oak tree" are bi-grams, and "big oak tree" is a tri-gram.

Natural Language Processing (NLP): a subfield of computer sciences primarily concerned with giving computers the ability to understand, interpret, and generate human language.

One-hot encoding: a technique used to represent categorical variables as numerical values in a machine learning model. It converts categorical variables into vectors of binary values, where each category is represented by a unique index in the vector.

One-shot learning: an extreme case of a few-shot learning problem, found mostly in computer vision, where a model is trained to recognize or classify new objects or patterns based on a single example.

Ontology: "A semantic model of the objects and their relationships in a domain of interest (Gruber, 1995); defines terms and concepts in a formal language that provides cross-references and semantic meaning; commonly published in OWL format for machine readability" from [88]

Out-of-vocabulary (OOV) words: words that are not present in the vocabulary or training set of a NLP or machine learning model. These are words that the model has not encountered during its training phase and, as a result, may not be able to properly handle or understand when encountered in new, unseen data.

Pipeline: in NLP, a series of processing steps involved in the analysis and transformation of raw text into a structured format. Each step in the pipeline performs a specific NLP task, such as tokenization, part-of-speech tagging, named entity recognition, and the output of one step becomes the input for the next.

Prompt: a specific instruction or query given to a language model or system to perform a task or generate a response.

Recommender system: a type of software application or algorithm designed to provide personalised suggestions or recommendations to users. The goal of a recommender system is to predict and suggest items that users might be interested in, based on their preferences, historical behaviour, or the behaviour of similar users.

Regular expression: a pattern of rules that can be used to search, match, or alter text. Regular expressions are capable of capturing variation in capitalisation, number and type of characters, adjacency of particular characters, which characters start or end the text, and wildcards allowing any characters in specific positions.

Self-attention: a weighing mechanism used in deep learning models, particularly in the Transformer architecture, which allows the model to focus on relevant parts of the input sequence when making predictions or generating outputs.

Sentence segmentation: a fundamental NLP process, also known as sentence boundary detection, which involves breaking down a given text into individual sentences.

Stemming: a text normalisation technique used in NLP to reduce words to their base or root form, known as the "stem." For example, the words “ecology”, “ecological” and “ecologist” can all be reduced down to the common word stem “ecology.”

Stop words: words which are commonly used in a language but which are generally considered to have little value in terms of meaning and are unlikely to contribute to the understanding of the content. These words are often filtered out in NLP to reduce the dimensionality of the data and focus on the most significant and informative words in a text.

Supervised learning: a machine learning paradigm in which a model is trained on a labelled dataset, i.e. a dataset where the input data (for example, a vector of predictor variables or an image) is paired with desired output labels or target values. The goal of supervised learning is to learn a mapping from input data to the corresponding output, enabling the model to make accurate predictions or classifications on new, unseen data.

Surface form: The form of a word as it appears in a text.

Self-supervised learning: a machine learning paradigm in which a model is trained using the data itself as a source of supervision, instead of relying on external labels provided by humans. This allows models to learn from large amounts of unlabelled data. Self-supervised learning has played a pivotal role in the emergence and success of large language models.

Term: a single word or a combination of words that convey a specific meaning. Terms are the building blocks of language, and in the context of NLP, they are often the units of interest when analysing and processing text.

Term-frequency-inverse document frequency (tf-idf): a numerical statistic used to evaluate the importance of a term (word) in a document relative to a collection of documents. The tf–idf is the product of two statistics, term frequency and inverse document frequency. Term frequency measures how often a term appears in a document. The inverse document frequency is a measure of how common or rare a term is across all documents. Higher tf-idf scores indicate that a term is both frequent within a specific document and rare across the entire corpus, making it more discriminative and potentially more important for representing the content of that document.

Tokenization & Tokens: Tokenization is the process of cutting up texts into small pieces (tokens) that can be processed by a computer. Historically text was segmented into “sentences” and “words” due to intuitive understanding of language and technical constraints. However, now more complex tokenization considering punctuation, word-forms, morphological derivatives, and sub-words are often used (see Mielke et al. (2021) for a survey of tokenization approaches from initial conception to the deep-learning era).

Topic modelling: a NLP technique used to identify topics present in a corpus. The goal is to automatically discover latent topics that characterise the main themes or subjects within a collection of documents.

Transformer: a type of deep neural network architecture. Transformer architecture is particularly well known for its ability to efficiently capture long-term dependencies in sequential data using the self-attention mechanism, making it well suited to tasks involving sequential data. Transformers have proven to be highly efficient and scalable, leading to the development of large pre-trained transformer models that have had a significant impact on various NLP tasks and beyond.

Unstructured data: any data that does not have a predefined data model or organisation. Unlike structured data, which is typically organised in rows and columns (e.g., relational databases), unstructured data lacks a clear and organised structure. Examples of unstructured data include text, multimedia data, web pages, sensor data, etc.

Unsupervised learning: a machine learning paradigm in which a model is trained on unlabeled data, and the goal is to discover the inherent structure or patterns present in the data without explicit supervision in the form of labelled outcomes.

Vector representation: the encoding of objects, such as words or documents, as vectors in a multi-dimensional space. Vector representations are used to represent the semantic meaning and relationships between objects in a way that can be leveraged by machine learning models. Two common types of vector representations are one-hot encoding and word embeddings.

Weak supervision: a scenario in machine learning where the training data is labelled with noisy, limited, or imprecise annotations instead of high-quality, fully accurate labels.

Word cloud: a visual depiction of the common words in a document, often with the size of each word in proportion to its relative frequency.

Word embedding: a dense vector representation of words in a continuous vector space such that words with similar meanings are closer to each other in the vector space. Word embeddings are classically learned or obtained from pre-trained language models and used as features in various NLP tasks.

Word sense: the various meanings or interpretations that a word can have in different contexts.