GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

Master's Thesis

Enhancing Ecological Knowledge Discovery Using Large Language Models

Author:	Viktor Domazetoski
Matriculation number:	29111214
Supervisors:	Dr. Patrick Weigelt, Dr. Terry Lima Ruas

Faculty of Forest Sciences and Forest Ecology
Göttingen, January 31, 2024

Contents

List of Figures

List of Tables

Acronyms

BOW Bag of Words. 28

CNN Convolutional Neural Network. 20

CV Computer Vision. 10

DL Deep Learning. 8, 11

GIFT Global Inventory of Floras and Traits. 4

GSC Gold Standard Corpus. 25

LLM Large Language Model. 12

LPD Living Planet Database. 17

LSTM Long Short-Term Memory. 11

ML Machine Learning. 10

NER Named Entity Recognition. 19

NLP Natural Language Processing. 8

NMAE Normalized Mean Absolute Error. 31

POWO Plants of the World Online. 23

QA Question Answering. 16

RNN Recurrent Neural Network. 11

ROUGE Recall-Oriented Understudy for Gisting Evaluation. 31

Abstract

The field of ecology is experiencing rapid growth, resulting in a surge of scientific literature, both contemporary and historical, spanning centuries. While this vast corpus of text holds a wealth of knowledge, the sheer volume makes it impossible for individuals to manually extract all the valuable insights it contains. Natural language processing (NLP) emerges as a powerful solution to tackle this challenge, offering a diverse array of applications. These applications range from classifying scientific papers to summarizing lengthy texts and even extracting structured data that can be employed in statistical models. NLP thus plays a pivotal role in unlocking the wealth of ecological knowledge buried within this extensive body of literature. In particular, Large language models (LLMs) have the power to revolutionize the field of ecological text analysis through their exceptional ability to comprehend, categorize, and extract valuable information from the copious amounts of ecological text data, and may enable researchers to navigate this knowledge-rich landscape with unprecedented efficiency and accuracy. To accomplish this objective, we evaluate a diverse set of encoder- and encoder-decoder-based LLMs across eight distinct tasks and languages, categorized into three domains: literature review, entity extraction and trait extraction. Within the literature review domain, our focus is on the topic modelling and text summarization tasks, enabling efficient data acquisition and processing from vast ecological text sources. In the realm of entity extraction, we employ named entity recognition models and family classification models to extract relevant entities from ecological texts. Lastly, we delve into the extraction of both categorical and numerical traits from ecological descriptions, encompassing text in English, Spanish, and German, while also exploring the model's performance in data-deficient scenarios. To facilitate model training and evaluation, we curate and utilize a range of datasets and gold standard corpora sourced from various related scientific papers. In summary, the LLMs displayed exceptional performance across all tasks and consistently outperformed the baseline models. Notably, in literature review-related tasks, the top-performing LLMs achieved F1-scores surpassing 88% when using paper titles and exceeding 95% when using abstracts. These models also excelled in text summarization, achieving ROUGE-L-SUM F1-scores exceeding 32% across the three datasets. For named-entity recognition, the best model achieved state-of-the-art F1-scores of 72.6% and 80.9% on two gold standard corpora. Furthermore, all LLMs achieved outstanding scores, consistently exceeding 95% in the family classification task. In the context of trait-related tasks, LLMs showcased their capability and versatility. achieving F1-scores ranging from 75% to 85% across English, Spanish and German descriptions. The few-shot learning approach demonstrated that LLMs can attain F1-scores exceeding 80% even with a training dataset as small as 128 labeled descriptions. Lastly, the models performed well in terms of normalized mean absolute error (NMAE), averaging 20.6%. Overall, these findings underscore the capabilities of LLMs in various ecological natural language processing tasks. The utilization of LLMs opens up new horizons for efficient knowledge extraction from the extensive body of ecological literature, offering unprecedented accuracy and productivity. As this field continues to evolve, LLMs promise to play an increasingly pivotal role in advancing ecological research and discovery, ultimately enhancing our understanding of the natural world and our ability to address pressing ecological challenges.

Zusammenfassung

Die ökologische Forschung erlebt ein rasantes Wachstum, was zu einer Flut an wissenschaftlicher Literatur sowohl zeitgenössischer als auch historischer Natur über Jahrhunderte hinweg führt. Obwohl dieser riesige Textkorpus eine Fülle von Wissen enthält, ist es aufgrund des schieren Umfangs für Einzelpersonen unmöglich, alle darin enthaltenen wertvollen Erkenntnisse manuell zu extrahieren. Die automatisierte Verarbeitung natürlicher Sprache (Natural Language Processing, NLP) erweist sich als leistungsstarke Lösung zur Bewältigung dieser Herausforderung und bietet eine Vielzahl von Anwendungen. Diese reichen von der Klassifizierung wissenschaftlicher Arbeiten über die Zusammenfassung langer Texte bis hin zur Extraktion strukturierter Daten, die in statistischen Modellen verwendet werden können. NLP spielt somit eine entscheidende Rolle bei der Erschließung des in dieser umfangreichen Literatur verborgenen Reichtums an ökologischem Wissen. Insbesondere Large Language Models (LLMs) haben das Potenzial, den Bereich der ökologischen Textanalyse zu revolutionieren, da sie über die außergewöhnliche Fähigkeit verfügen, die umfangreichen Mengen ökologischer Textdaten zu verstehen, zu kategorisieren und wertvolle Informationen daraus zu extrahieren, und es Forschern ermöglichen, sich mit beispielloser Effizienz und Genauigkeit in diesem Wissen zurechtzufinden. Um dieses Ziel zu erreichen, evaluieren wir einen vielfältigen Satz von Encoder- und Encoder-Decoder-basierten LLMs für acht verschiedene Aufgaben, die in drei Bereiche kategorisiert sind: Literaturrecherche, Entitätsextraktion und Merkmalsextraktion. Im Bereich der Literaturrecherche liegt unser Fokus auf Themenmodellierung und Textzusammenfassungsaufgaben, die eine effiziente Datenerfassung und -verarbeitung aus umfangreichen ökologischen Textquellen ermöglichen. Im Bereich der Entitätsextraktion verwenden wir Modelle zur Erkennung benannter Entitäten und Familienklassifizierungsmodelle, um relevante Entitäten aus ökologischen Texten zu extrahieren. Abschließend beschäftigen wir uns mit der Extraktion sowohl kategorialer als auch numerischer Merkmale aus ökologischen Beschreibungen, einschließlich Texten in Englisch, Spanisch und Deutsch, und untersuchen gleichzeitig die Leistung des Modells in Szenarien mit Datenmangel. Um das Training und die Bewertung von Modellen zu erleichtern, kuratieren wir eine Reihe von Datensätzen und stellen Goldstandard-Korpora zusammen, die aus verschiedenen verwandten wissenschaftlichen Arbeiten stammen. Zusammenfassend lässt sich sagen, dass die LLMs bei allen Aufgaben eine außergewöhnliche Leistung zeigten und die Basismodelle durchweg übertrafen. Bemerkenswert ist, dass die leistungsstärksten LLMs bei Aufgaben im Zusammenhang mit der Literaturrecherche bemerkenswerte F1-Werte erzielten, die bei der Verwendung von Titel von Veröffentlichungen über 88% und bei der Verwendung von Abstracts über 95% lagen. Diese Modelle zeichneten sich auch bei der Textzusammenfassung aus und erzielten in allen drei Datensätzen beeindruckende ROUGE-L-SUM F1-Werte von über 32%. Bei der Entitätsextraktion erzielte das beste Modell besonders hohe F1-Werte von 72,6% und 80,9% bei zwei Goldstandard-Korpora. Darüber hinaus erzielten alle LLMs hervorragende Ergebnisse und lagen bei der Familienklassifizierungsaufgabe durchweg über 95%. Im Kontext merkmalsbezogener Aufgaben stellten LLMs ihre Leistungsfähigkeit und Vielseitigkeit unter Beweis und erreichten von F1-Ergebnisse zwischen 75% und 85% bei englischen, spanischen und deutschen Beschreibungen. Der Few-Shot-Learning-Ansatz hat gezeigt, dass LLMs selbst mit einem Trainingsdatensatz von nur 128 beschrifteten Beschreibungen

F1-Scores von über 80% erreichen können. Schließlich schnitten die Modelle hinsichtlich des normalisierten mittleren absoluten Fehlers (NMAE) mit durchschnittlich 20,6% hervorragend ab. Insgesamt unterstreichen diese Ergebnisse die bemerkenswerten Fähigkeiten von LLMs bei verschiedenen ökologischen Aufgaben der Verarbeitung natürlicher Sprache. Der Einsatz von LLMs eröffnet neue Horizonte für die effiziente Wissensextraktion aus der umfangreichen ökologischen Literatur und bietet beispiellose Genauigkeit und Produktivität. Da sich dieser Bereich ökologisch weiterentwickelt, versprechen LLMs, eine immer wichtigere Rolle bei der Weiterentwicklung der ökologischen Forschung und Entdeckung zu spielen und letztendlich unser Verständnis der natürlichen Welt und unsere Fähigkeit, drängende Herausforderungen anzugehen, zu verbessern.

1 Introduction

Earth is home to an astonishing diversity of life forms. Among these, vascular plants stand as one of the most vital and ubiquitous groups, accounting for approximately 80% of global biomass [3]. The census of vascular plants, which presently exceeds 380,000 identified species globally [45, 96, 76], is evidence to the remarkable diversity within this botanical realm. Furthermore, this number may be an underestimate given that, during the past few decades, about 2,000 new species have been described per year [17]. Vascular plants play a critical role in driving crucial ecological processes such as carbon sequestration, nitrogen cycling, and habitat provision because they are the principal producers in ecosystems. Their profound ecological significance cannot be overstated, as they are the foundation of complex food webs, provide sustenance for myriad herbivores and thus exert an overarching influence on trophic dynamics. Beyond their ecological roles, vascular plants have far-reaching socio-economic implications, as they serve as sources of food, medicine, and raw materials, with cultures around the world relying on them for sustenance, cultural practices, and traditional medicine.

Despite the large number of species, a substantial portion of our knowledge regarding these organisms remains incomplete [23]. Out of the estimated 380,000 species, many lack comprehensive taxonomic descriptions, and even more have not been subject to in-depth scientific investigation. This is particularly evident in regards to crucial aspects such as their geographical distribution, functional traits, and ecological contributions [138, 83]. This knowledge gap arises from various factors, including limited funding for botanical research, inadequate access to remote regions, and the sheer scale of plant diversity. Nonetheless, knowledge is a necessity to understand the role within their ecosystem [133] and their contribution to ecosystem services [97], such as carbon sequestration, soil stabilization, and water purification. Our ability to conserve and protect threatened plant species is hampered when we lack basic information about their biology, distribution, and ecological role. This results in a significant shortage of data to anticipate which areas are best to focus on in the conservation of these species. Especially in the age of the climate crisis [90, 31] and on the brink of the sixth mass extinction [25, 8], it is crucial to understand how different plant species respond to shifting environmental [52] and anthropogenic [123] factors to be able to make reliable forecasts on the state of these plants.

Ecological databases hold information on a wide range of plant species, including measurements on physiological, chemical and genomic traits at the individual or species level, data on species' distribution and invasiveness, and have revolutionized ecological research in several profound ways. There are several databases that hold such information, including the Global Biodiversity Information Facility (GBIF) [1], the Global Inventory of Floras and Traits (GIFT) [131], the Global Invasive Species Database (GISD) [43], the Botanical Information and Ecology Network database (BIEN) [79], the Open Traits Network [39], the Austraits database [33], and the TRY Plant Trait Database [58, 57]. These databases make them essential resources for a broad spectrum of ecological fields and across scales [60] and allow the study of ecology as a big-data science [34]. All of this data led to a vast number of studies in the field of functional ecology and advancing our comprehension of

[1] GBIF.org (2023) Available at: https://www.gbif.org

the worldwide patterns in plant form and function [29, 85, 137].

One of the most significant contributions of these databases to ecological research is the democratization of data. GBIF, for instance, aggregates biodiversity data from across the globe, and makes it freely accessible to researchers, decision-makers, and the general public. It promotes transparency and enables researchers to expand on existing knowledge, accelerating the pace of ecological discovery. This abundance of data makes it possible for scientists to examine patterns and trends in biodiversity at unprecedented scales, providing a comprehensive understanding of how ecosystems function and evolve over time. It is also instrumental in understanding the impact of climate change [64], habitat loss, and invasive species on ecosystems [56, 38]. Moreover, researchers can access raw data, methodologies, and metadata, ensuring the transparency and reproducibility of scientific studies [93]. By basing their judgements on current, reliable ecological data, policymakers and land managers can improve the sustainability of current natural resource management practices.

Furthermore, these databases foster interdisciplinary research by integrating diverse datasets. Researchers from diverse backgrounds can access and contribute to these databases, fostering global collaboration in ecological research. This cooperative strategy encourages the exchange of knowledge, data, and methodologies across scientific communities, ultimately leading to more thorough and robust ecological studies. Ecological research often requires the collaboration of experts from various fields, including botany, zoology, climatology, and geology. These researchers may look at how species interact with their environment by utilizing databases like GIFT and TRY, which contain a comprehensive inventory of plant traits[59]. By integrating these traits with environmental data or alien and invasive species data[43], they can gain further insights into ecosystem dynamics, species distributions, and responses to environmental changes[7, 114]. Ecological databases also have an ability to support conservation efforts, such as assisting in the identification of high biodiversity areas and helping conservationists prioritize regions for protection and restoration [7]. These databases can also be instrumental in monitoring biodiversity changes over time, facilitating the early detection of threats such as habitat loss, invasive species, and the effects of climate change [94]. In this way, they contribute to the creation of evidence-based conservation strategies and policy development [126].

While ecological databases such as GBIF, GIFT, and TRY have undoubtedly transformed ecological research, they are not without their limitations and biases. First of all, although databases such as TRY host a vast trove of plant data, a much larger portion is missing. Regrettably, out of the over 380,000 globally identified plant species, only a fraction, approximately 130,000, have records that contain information on at least one out of the 1,918 distinct traits contained in TRY (Fig.1a). Out of these, 97,439 possess limited trait data, encompassing fewer than ten traits. Only 6,331 species, or 2% of all plant species, contain information on more than 50 traits. While GIFT has a larger overall coverage of traits, similar patterns emerge. In GIFT, approximately 290,000 (77% of all plant species) contain at least some trait information and 139,897 species contain information on at least 10 traits out of the 109 contained in GIFT. However, only 939 species contain measurements on more than 50 traits. The traits themselves have a significant impact on their coverage. This depends on the usefulness of the trait as well as how easily the trait is measured [80], leading to some traits such as the plant's growth form having much higher coverage than other, arguably more important traits such as the plant's maximum height or specific leaf area. In the TRY database, out of the 1,918 measured traits, 1,345 contain fewer than 100 measurements, 221 contain between 1,000 and 9,999 measurements and only 24 are measured for more than 10,000 species (Fig.1b). This gap is less pronounced in the GIFT dataset's 109 measured traits, as 9 traits contain fewer than 100 measurements and 79 traits contain more than 1,000 measurements. However, out of these 79 traits, only 37 traits are measured for more than 10,000 species. Overall, the mean global completeness across all traits in the TRY dataset is only 0.21% and the median trait completeness is 0.0051% [80], while the mean global completeness across all traits in the GIFT dataset is 7.25% and the median trait completeness is 1.1784%. This dearth of comprehensive trait information for the majority of plant species hampers our capacity to conduct trait-based ecological analyses, constrains our understanding of their functional roles in ecosystems, and impedes efforts to predict their responses to environmental changes.

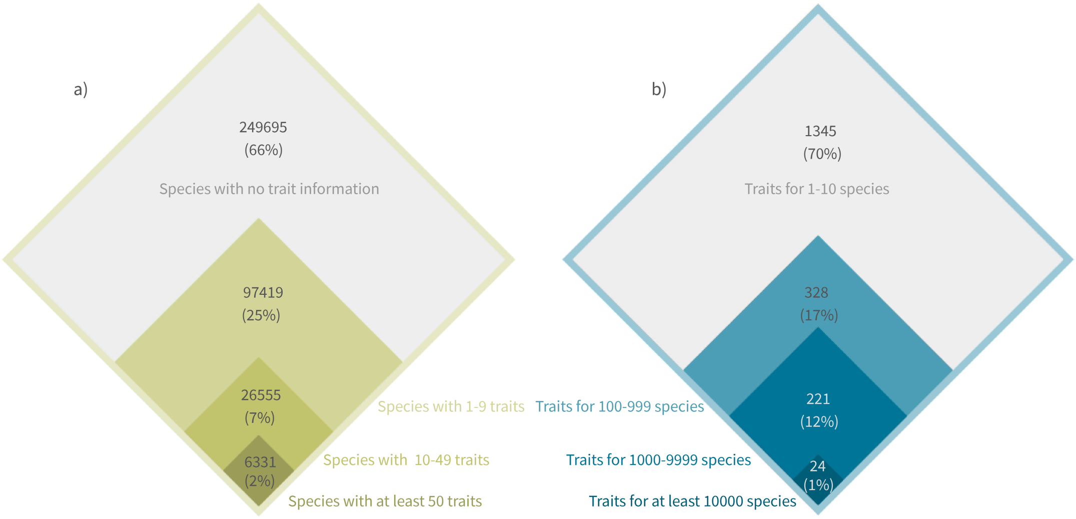

Figure 1: The number (percentage) of plants for which there are none, between 1-9, between 10-49 and at least 50 traits in TRY (a), and the number (percentage) of traits for which there are between 1-99, between 100-999, between 1000-9999 and at least 10000 species in TRY (b). Created using TRY data taken from [80].

Secondly, a prominent concern in these databases is the existence of geographical and socioeconomic biases [4, 80, 95], which can significantly impact the quality and scope of ecological studies. Geographical biases in ecological databases are particularly noteworthy. An illustrative example can be drawn from the TRY dataset, where, for a subset of 53 traits, trait coverage exhibited a considerable range, spanning from 1.01% to 32% [80]. Furthermore, the mean completeness across traits ranged between 2.8% in New Guinea to and 58.7% in the Faroe Islands. These biases arise because data collection efforts tend to be more extensive in certain regions, such as North America and Europe, compared to less-studied areas like parts of Africa, South America, and Asia. Consequently, data points are distributed unevenly, disproportionately favoring regions of high biodiversity, well-developed research infrastructure, and economic resources. As a result, the data available in these databases may not accurately reflect the true global distribution of species and ecosystems. Researchers that primarily rely on these databases may inadvertently reinforce existing knowledge gaps, potentially overlook critical conservation needs in these areas, and therefore hinder our understanding of biodiversity in underrepresented regions. Addressing this bias requires targeted efforts to enhance data collection and accessibility in underrepresented regions, fostering a more holistic understanding of global ecological dynamics.

Socioeconomic biases are another related challenge. Ecological data collection requires resources, including funding, trained personnel, and access to technology. Using a set of 53 focus traits from TRY, [80] found four significant predictor variables of trait completeness. Out of these, the mean species range size and research expenditure were positively correlated to the completeness of traits, while plant endemism and species richness were negatively correlated. Consequently, research efforts are often concentrated in wealthier nations and institutions, while many developing countries and less-privileged communities have limited access to resources for ecological research and data sharing. This disparity can lead to an underrepresentation of data from marginalized regions and communities. The incompleteness of ecological databases can introduce uncertainties and biases into research results. Many areas, particularly remote or politically unstable regions, lack comprehensive biodiversity data. This data gap can hinder our ability to assess global biodiversity accurately and make informed conservation decisions. The quality of data in ecological databases can also vary significantly, with entries being outdated, misidentified, or lacking essential metadata. Researchers must exercise caution and rigor when using these databases to avoid propagating inaccuracies and biases in their analyses. Research activities can also be biased toward certain taxa or ecosystems. Often, charismatic or economically valuable species receive more attention, while less conspicuous or commercially unimportant species remain poorly understood. This bias may not properly account for the ecological significance of less well-studied species, thus skewing conservation priorities and potentially even distorting the broader ecological picture.

Finally, the reliance on existing data can inadvertently promote biases in research results [4]. Researchers may choose to work with readily available data rather than investing in new data collection efforts, which can perpetuate the aforementioned biases present in the databases. This can have implications for the accuracy of ecological models, conservation prioritization, and policy recommendations. Efforts to promote inclusivity in data collection, data-sharing partnerships with underrepresented regions, and a critical awareness of these biases are essential steps towards improving the accuracy and relevance of ecological research worldwide.

Hidden within the pages of historical texts, national floras, and inventories, as well as more recent publications, lies a treasure trove of ecological information that remains largely untapped by existing ecological databases. While databases like GIFT have made significant strides in aggregating some of this ecological data, a vast reservoir of knowledge still resides outside their digital confines. Mobilizing this wealth of hidden information could significantly expand the coverage and depth of our ecological understanding. Historical texts are a goldmine of ecological insights, often documenting species distributions, ecological observations, and environmental conditions dating back centuries. Naturalists and explorers from different eras have left behind invaluable records of ecosystems, species interactions, and environmental changes. The Biodiversity Heritage Library (BHL, [48]) launched in 2009 and estimates that there are more than 120 million pages published in over 5.4 million books since 1469, plus about 800,000 monographs and 40,000 journal titles. Fifty percent of these were published before 1923 and are in the public domain in the United States. As of the 12th September, 2023, the BHL [2] holds information on 188,436 titles, 300,344 volumes and 61,231,984 pages. By digitizing and integrating these historical records into ecological databases, researchers can unlock a plethora of information. This in turn will enable them to examine long-term trends, shifts in species distributions, and responses to environmental changes.

National floras and inventories provide comprehensive documentation of plant species within particular geographic areas. These authoritative references offer detailed descriptions, distribution maps, and ecological information for numerous plant species[131]. By incorporating this rich resource into ecological databases, we can better comprehend regional biodiversity patterns, identify knowledge gaps, and inform conservation efforts. Newer publications, especially those emerging from biodiversity surveys and ecological studies in less-explored regions, harbor vital data that can complement existing databases. These studies often unveil previously unknown species, document rare and endangered species, and provide context for ecological dynamics in specific ecosystems. By actively integrating data from recent research into ecological databases, we can ensure that the most current and relevant information informs scientific inquiry and conservation decision-making. The field of ecology is producing an increasing volume of scientific literature, with over more than 80,000 articles published between 1980 and 2019 [81]. Mobilizing these untapped sources of ecological information demands collaborative efforts among researchers, institutions, and data repositories. Digitization projects [108, 87], citizen science initiatives [37], and international partnerships [121] can facilitate the process of extracting, standardizing, and disseminating this invaluable ecological knowledge. Embracing the diversity of information sources, from historical manuscripts to contemporary fieldwork, not only expands the coverage of ecological databases but also enriches our understanding of the intricate web of life on Earth. Ultimately, unlocking the hidden ecological insights within these texts promises to fuel more comprehensive and informed ecological research.

Natural Language Processing (NLP) holds tremendous promise for ecologists as it allows for the extraction of information from large volumes of scientific literature, which can provide valuable insights into the relationships between organisms and their environment [35]. The use of NLP can aid efficiently process the vast amount of historical and published research on ecology, helping ecologists identify patterns, trends, and knowledge gaps that may not be immediately apparent through traditional methods of literature review. The area of NLP itself is a rapidly growing field with many approaches applicable to ecology. Prior to a few years ago, NLP algorithms were mostly utilised for a few more straightforward tasks such as the classification of text for tasks like sentiment analysis, and the named entity recognition of text [53]. However, with the rise of performance and adaptability of deep learning (DL) models [75], other more complex tasks such as summarization, question answering and text generation can be employed to extract valuable information from text.

[2] https://www.biodiversitylibrary.org/

Ecologists can benefit from NLP in a number of ways, including:

- **Literature reviews and knowledge discovery**: Conducting literature reviews is a fundamental step in ecological research, but it can be time-consuming and challenging due to the sheer volume of available literature. Text mining techniques can streamline this process by automatically identifying and categorizing relevant articles, extracting key information, and summarizing research findings [71]. This enables ecologists to stay up-to-date with the latest research and unearth insightful information that has been hidden within a vast body of literature.

- **Ecological pattern detection**: Text mining can reveal ecological patterns and trends by analyzing massive text corpora [26, 24]. For instance, it can identify emerging topics or research gaps in the body of ecological literature, assisting researchers in the prioritization of areas for further investigation. By examining reports and observational data, it can also be used to detect shifts in species ranges, the impacts of climate change, or the spread of invasive species.

- **Data mobilization and synthesis**: Text mining can facilitate the mobilization of ecological data by automating the extraction of relevant information from a wide range of sources [65, 30]. It enables ecologists to efficiently gather data on species distributions, ecological interactions, functional traits, environmental conditions, and more from scientific literature and reports. Text mining accelerates the data synthesis process by condensing and structuring this information, making it easier for researchers to create extensive datasets that can be used for further analysis. Additionally, text mining can contribute to data quality assurance and standardization [1], by identifying inconsistencies, errors, or ambiguities in textual data. This would help researchers to guarantee that the information they collect is reliable and consistent, which is a crucial aspect for robust ecological analyses and modeling.

- **Multilingual insights**: Ecological research is often global in scope and involves multilingual literature. Exclusion of non-English literature may further bias studies and meta-analyses [62]. The advancement of open-access multilingual NLP models such as BLOOM[103] provides a innovative approach to process and analyze text in numerous languages, breaking down language barriers and facilitating the integration of data from diverse sources [6, 2].

However, in comparison to certain other scientific disciplines, ecologists have not fully harnessed the power of NLP. While NLP has become increasingly prevalent in fields such as biomedicine or economics [35], where it is used for tasks such as text mining of medical literature and electronic health records, its integration into ecological research has been more limited.

One factor for the relatively slow adoption of NLP in ecology may be the traditional emphasis on fieldwork, data collection, and statistical analysis. Ecological research has historically prioritized direct observations and experiments, which has resulted in a reliance on quantitative data. However, as the volume of ecological literature and textual data grows and as the digitization of historical texts allows for global access to these texts, there is a growing recognition of the necessity for advanced text analysis techniques offered by NLP.

It is interesting to note that some ecological subfields have seen greater collaboration between ecologists and computer science experts such as in bioacoustics and computer vision. In bioacoustics, for instance, machine learning (ML) and NLP are employed to analyze and interpret animal vocalizations and sounds in natural settings [112]. Similarly, computer vision (CV) techniques are applied to process imagery and remotely sensed data for ecological purposes, such as species identification and habitat monitoring [111, 119]. Machine and deep learning methods have also been used in more traditional ecological tasks [92, 18], such as species distribution modeling [22] and species richness models [7]. These collaborations have yielded cutting-edge tools and methodologies that enhance ecological research by leveraging advances in technology and data analysis.

Within the thesis, we analyze the potential of large language models in a variety of tasks. We start by considering two NLP tasks aimed at enhancing the efficiency of reviewing ecological papers. These tasks include the summarization of articles and abstracts, offering a proxy to efficiently assess the content of the papers by condensing it to its most pertinent details. Furthermore, we incorporate a topic modeling task, in which we categorize papers based on their relevance to two extensive macroecological databases. Next, we delve into two tasks centered around the extraction and categorization of entities. We do this is by evaluating the large language models on a named entity recognition task. Additionally, we evaluate the models' proficiency in predicting the family of a species given its description. This task is useful for the categorization of descriptions, and highlights the models' ability to independently acquire rules such as those found in plant identification keys. The final tasks we consider relate to the extraction of trait information embedded within species' descriptions. We first start with the extraction of categorical traits, framing this task as a sequence classification problem. We extend this task to encompass Spanish and German descriptions by leveraging multilingual and other language-specific monolingual models. To investigate the data deficiency problem which is present for a large number of traits, we also analyze how model performance changes with gradually increasing subsets of data. To address this issue, we propose a few-shot learning approach as a potential solution. Finally, we demonstrate the applicability of extractive question answering models in extracting numerical traits from textual description, offering a holistic perspective on trait information extraction.

The thesis is organized in the following manner: Firstly, we give an introduction to natural language processing, deep learning and large language models. We give information on the components of large language models, the diverse architectures and the spectrum of natural language processing tasks they are tailored for and that we will use in this work. In section 3, we formulate all of the eight distinct tasks under three overarching sections: literature review, entity extraction and trait extraction. Section 4 is dedicated to elucidating the datasets employed to train and evaluate the models, emphasizing the preprocessing steps essential to fit them to the model's input requirements. Section 5 outlines the adopted methods, encompassing the various language models utilized, our approach to task evaluation and in-depth insights into model implementation. The outcomes of each task are shown in section 6, followed by a comprehensive discussion on the current state of large language models, their potential, and limitations in section 7. The concluding section summarizes our work and offers a glimpse into potential avenues for future research.

2 Research area introduction

2.1 Natural language processing

Natural language processing (NLP) represents a cutting-edge technological field that acts as the vital link between human language and computer comprehension. It operates at the intersection of diverse disciplines, including linguistics, artificial intelligence, and computer science, with the primary aim of enabling machines to engage with human language in a meaningful manner [16]. NLP has far-reaching effects, revolutionizing fields like healthcare, finance, customer service, and beyond by facilitating effective human-computer communication and automating various linguistic tasks. At its core, NLP seeks to equip computers with the capability to interpret and process natural language, which is fundamentally complicated and context-dependent. This encompasses a broad spectrum of tasks, ranging from fundamental functions like text categorization and sentiment analysis to more intricate endeavors, including machine translation, speech recognition, and question-answering systems. NLP systems are meticulously designed not only to grasp the surface-level syntax and semantics of language but also to comprehend the nuanced intricacies, idiomatic expressions, and cultural diversities that hallmark human communication. The significance of NLP becomes increasingly evident in our digitized world. With the exponential surge in textual data across the internet, social media, and digital communication platforms, NLP plays an indispensable role in extracting valuable insights from this vast information pool. Furthermore, there are numerous practical applications of NLP, with virtual assistants like Siri and Alexa, chatbots employed in customer support, and healthcare systems that use it to analyze medical records and aid in diagnoses.

2.2 Deep learning

Deep learning (DL) has brought about a paradigm shift in the realm of NLP, introducing more potent and efficient techniques for handling and analyzing extensive textual data [117]. One of the initial uses of DL within NLP was the introduction of word2vec [84], which employs neural networks to create embeddings of words. These word embeddings serve as vector-based representations of words in a spatial framework, aiming to encapsulate both their semantic and syntactic nuances. Given that all statistical models, including deep learning models, necessitate numerical input, the word embedding has become the primary means of translating textual data into meaningful numerical representations and analogous embeddings have been used to embed information from different sources in fields like bioacoustics and network science.

Since the inception of word2vec, there have been notable advancements in the domain of deep learning models tailored for NLP. An example of such progression is the emergence of recurrent neural networks (RNN) and their variants, including long short-term memory (LSTM) networks [107]. Initially designed to cater to time series data due to their ability to retain information from preceding predictions, RNNs exhibited the capacity to process textual sequences, enabling them to apprehend the temporal interdependencies among words within a sentence or document. However, these architectures were ultimately impeded by the vanishing gradients problem, which pertains to the issue of gradients becoming exceedingly small as they are propagated backwards through the layers of the network during training. When the gradients vanish, it means that the updates applied to the network's weights during the optimization process become negligible, making it difficult for the model to learn and capture long-range dependencies in the data. While LSTMs' gating mechanisms helped mitigate the this to some extent, the problem persisted.

The pinnacle advancement in deep learning for NLP materialized with the advent of the transformer architecture[122]. Transformers employ self-attention mechanisms for text processing, affording them the capability to capture overarching relationships between words within a document. Self-attention allowed the model to assign different levels of importance to different parts of the input sentence, enabling it to capture long-range dependencies more effectively and bringing forward a more comprehensive solution to the vanishing gradients problem. This attribute renders them highly efficacious for a spectrum of tasks, including language modeling, named entity recognition, and text classification. The transformer models steered the research direction of the field toward foundation models, characterized by billions of adaptable parameters. These models harness the concept of transfer learning, where the core of the model comprehends the language's vocabulary and word connections. This core language model is subsequently coupled with specialized "heads" fine-tuned for specific tasks, such as classification, summarization, or question answering.

2.3 Large language models

Transformers, as foundational models, revolutionized not only the field of NLP, but also extended their transformative impact to domains such as computer vision (CV) and digital signal processing, sparking a paradigm shift in the construction of DL models. Central to this transformation is the concept of transfer learning, which posits that knowledge acquired from one task or dataset can be successfully applied to another with a similar context, forming the bedrock of the foundational model paradigm.

The integration of the attention mechanism [122] within transformers marked a pivotal advancement, enabling the model to grasp long-range dependencies in data without relying on recurrent layers, as was the case in earlier models. This mechanism introduced the ability to learn contextually, allowing word representations in vector space to dynamically adapt in response to their contextual surroundings. Consequently, words like "bank" could now be meaningfully represented, taking into account their associations with financial institutions or riverbanks, a capability that was notably absent in earlier models. This contextual understanding was crucial in catapulting so called large language models (LLM) to a state-of-the-art level of performance.

The architecture of transformers can encompass one or both of the following components: an encoder and a decoder. This flexibility empowers the model to excel in a diverse array of tasks, making it a versatile choice for various applications in NLP and beyond.

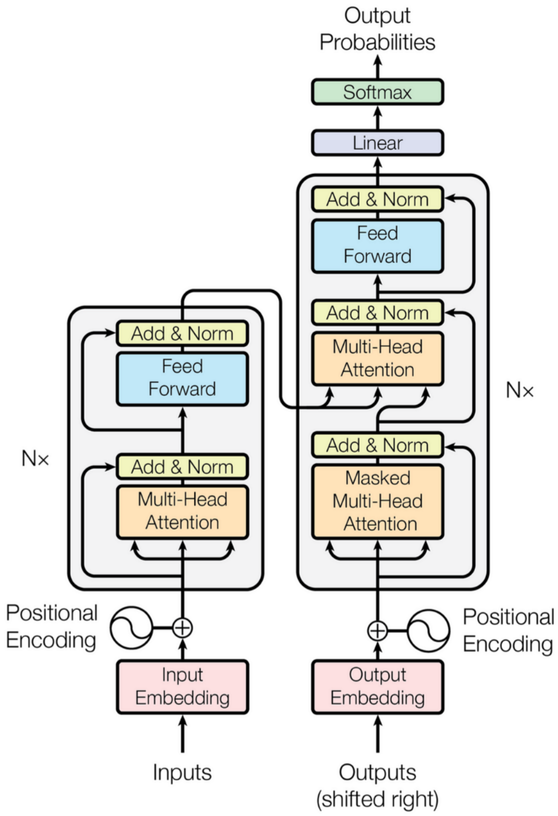

Figure 2: The original transformer architecture. The model consists of an encoder (left) and a decoder (right), both of which contain the attention mechanism (orange). Taken from [122].

- **Encoder**: often called an *auto-encoding* model, is a type of LLM that excels at understanding and encoding input data into a fixed-length representation known as a context vector or embedding. It processes sequential or unstructured data, such as sentences or documents, and transforms them into a numerical format that captures the semantic information and context, using a so called "bi-directional" attention mechanism. This indicates that at each stage, the attention layers can access all words within the input sentence. The representation is then used as input for various downstream tasks. This unique feature equips them with a notable advantage in tasks that demand a holistic understanding of the entire textual input, such as sequence classification, sentiment analysis, named entity recognition, and extractive question answering. Popular encoder models include the Bidirectional Encoder Representations from Transformers (BERT) [28], its variants like RoBERTa, DeBERTa [72, 51], and others like ELECTRA [21].

- **Decoder**: frequently referred to as an *auto-regressive* model, specializes in generating sequences of data, such as sentences or paragraphs, based on an input context. Consequently, decoders are most often used in applications like text generation, language translation, and chatbots. Some of the most well-known decoder models fall in the Generative Pre-trained Transformer (GPT) series [98], as well as the more recent, LLaMa architecture [118]. These models employ auto-regressive generation techniques, meaning that the attention layers can exclusively access words preceding the current position. The decoder then predicts one token at a time while conditioning on previously generated tokens. Using this tactic and by employing a probabilistic language model, decoders can generate coherent and contextually relevant text based on a given input.

- **Encoder-decoder**: or *sequence-to-sequence*, architectures combine the strengths of both encoders and decoders to perform tasks involving translation, summarization, question-answering, and more. In this architecture, the encoder processes the input data and generates a context vector, which is then passed to the decoder to produce the desired output sequence. For instance, in machine translation, the encoder encodes the source language sentence, and the decoder generates the equivalent sentence in the target language. Popular encoder-decoder architectures include Google's T5 and its FLAN variant [99, 19], and Meta's BART [67]. These models have been highly successful in various natural language processing applications and have achieved state-of-the-art results in machine translation and text summarization.

All transformer models, regardless of whether they function as encoders, decoders, or encoder-decoders, share a common structure comprising three essential elements: the tokenizer, the language model, and the task-specific head.

2.3.1 Tokenization

The tokenizer serves a critical role in the preprocessing of raw input text. Its main purpose is to segment the continuous input text into discrete units known as tokens, which are subsequently processed by the model. In essence, the tokenizer acquaints itself with the model's vocabulary and converts the input text into a format that can be comprehended and manipulated by the machine. Tokenization encompasses the act of segregating words, punctuation, and other textual elements into these tokens, essentially translating the linguistic content into machine-readable data. Tokenizers vary in their approaches, with the three fundamental categories being: word-based, character-based, and subword-based. Word-based tokenizers operate by considering entire words within their designated vocabulary. While this approach allows them to accommodate a vast array of words, it comes at the cost of an expansive vocabulary, potentially consuming substantial memory resources. This is especially apparent when considering the extensive lexicon of languages like English, which comprises over 500,000 unique words. Furthermore, word-based tokenizers face a limitation: words unseen during the tokenizer's training are represented by an unknown token, thereby curtailing the potential for effective transfer learning. On the contrary, character-based tokenizers [127, 98] utilize a significantly smaller vocabulary, but require that each word be represented as a combination of character tokens, which can impact the model's performance. Finally, subword tokenization algoriths [106, 63] combine the advantages of both word-based and character-based tokenization. They operate under the principle that frequently used words should be learned as whole units, while less common terms should be deconstructed into meaningful subword components. This dynamic approach strikes a balance, offering a practical and memory-efficient solution that enhances the model's ability to handle a broad spectrum of linguistic data while mitigating vocabulary size concerns and promoting effective transfer learning.

The vocabulary of the tokenizer is the base of all transformer models and is dependent on the training corpus of the model. BERT based models are usually trained on the English Wikipedia Corpus which comprises 2.5 billion words and the BookCorpus which comprises 0.8 billion words. Both of these sources contain information that is generic in nature and not particularly tailored to any one scientific field. To investigate whether a relevant scientific vocabulary may improve task performance, studies have explored the avenue of training models from scratch specifically to adapt the vocabulary to a specialized domain. One notable example is SciBERT [5] which has its own SciVOCAB as it is trained on 1.14M papers from Semantic Scholar. PubMedBERT [46] is similarly is trained on approximately 3.1 billion tokens from PubMed full text articles and therefore creates its own biomedical vocabulary.

2.3.2 Language model & pre-training

The primary objective of the language model within the transformer architecture is to generate a text representation that can be used as input by the model's task-specific head. This process can be achieved by training the model in several ways. Typically, encoder models are guided by two fundamental objectives: masked language modeling and next-sentence prediction, while decoder models are trained using a causal masked language modeling objective. Masked language modeling entails the task of predicting masked or hidden words within a given sentence. This process encourages the model to grasp the meanings of individual words and their contextual usage within a sentence. Concurrently, next-sentence prediction requires the model to predict the subsequent sentence in a given sequence of text. Through this task, the model learns to understand the logical flow and coherence between sentences, enhancing its ability to generate contextually relevant text. As a result of these training objectives, words with intrinsic semantic connections, such as "Japan" and "sushi," are positioned closer to each other in the embedding space, reflecting their natural associations. Conversely, words with dissimilar or unrelated meanings, like "Japan" and "pizza," are positioned further apart within the embedding space, thus illustrating the model's capacity to capture semantic relationships and contextual nuances in text data [84]. Pre-training the language model has proven to be an efficient method [47] and is the approach of models such as BioBERT [66] in the biomedical domain and FinBERT [73] in the financial domain. Pre-training like this is done through the use of the above-mentioned objectives and by minimizing a metric called perplexity which indicates how much the model is "perplexed" by unseen examples and suggests it has learned the basic patterns of grammar of the language.

2.3.3 Task-specific head & fine-tuning

The task-specific head, positioned atop the pre-trained language model, represents the final component of the transformer architecture. In a crucial step known as fine-tuning, the transformer's weights are immobilized, and the task-specific head undergoes training using the embedding produced from the language model as input and dedicated task-specific data as output. Fine-tuning is a prevalent approach in transfer learning, serving as a mechanism to acquire task-specific expertise to the model while preserving the broader knowledge acquired during pre-training. The specifications of the head vary for each specific task. We outline these details for the tasks within the scope of this work:

- **Sequence classification**: A sequence classification head is a neural network component designed to categorize the word embedding sequences generated from the language models into predefined classes or categories. It typically operates by processing the final hidden state of the input sequence, often the [CLS] token embedding, through additional layers such as fully connected or softmax layers. This enables the model to learn patterns and features within the input sequence that are indicative of the target class, making it well-suited for tasks like sentiment analysis, text classification, and categorical trait prediction. For our scenario, sequence classification will be used in several tasks, including the modeling of a paper's relevance, the classification of a species' taxonomic family based on a description, and the extraction of categorical functional traits from English, Spanish and German descriptions.

- **Token classification**: A token classification head is a component used for sequence labeling tasks, where the goal is to assign labels or categories to individual tokens within a sequence. It operates by processing the embeddings of each token and predicting the label for each token separately. This makes the model useful for tasks such as named entity recognition, part-of-speech tagging, and semantic role labeling.

- **Extractive question answering**: An extractive question answering (QA) head is designed for the task of extracting answers directly from a given text in response to a question. It works by identifying specific spans of text within the input document that contain the correct answer to the posed question, unlike generative QA, which provides an answer by generating new text. Extractive QA heads often employ techniques like token-level classification or pointer networks to pinpoint the precise sections of the text that correspond to the answer. This approach is particularly useful for tasks where the answers are present in the text and don't require generation, such as fact-based question answering or document summarization.

- **Text summarization**: A text summarization head is a crucial component within a neural network designed for the task of condensing lengthy textual content into concise and coherent summaries. It functions by extracting the most salient information from the input text and crafting a shorter version that retains the key points and essential details. Text summarization heads can employ methods such as attention mechanisms and generation networks to identify significant sentences or phrases and construct a summary that conveys the main ideas present in the original text. This capability makes them invaluable for tasks like document summarization, news article abstraction, and content compression.

- **Few shot learning**: A few-shot Learning head is a component within a machine learning model that specializes in tasks requiring the model to make predictions based on a limited amount of labeled data, typically referred to as "few shots." It operates by adapting the model's parameters based on the available few-shot examples and their corresponding labels. This enables the model to generalize from the provided data and make accurate predictions on unseen samples, making Few-shot Learning heads well-suited for scenarios where acquiring extensive labeled data is impractical.

3 Task formulation

In this thesis, we will frame the language modelling tasks as part of the NLP paradigms mentioned above. We will commence by addressing tasks associated with literature review improvement and information extraction enhancement. Subsequently, we will demonstrate the application of NLP techniques for entity extraction from textual data. Finally, we will explore methods for extracting trait-related information in a diverse set of scenarios.

3.1 Literature review

We begin by addressing two tasks aimed at automating the extraction of relevant information from ecological literature: topic modeling and text summarization.

3.1.1 Topic modelling

The task of relevant paper identification can be effectively formulated as a sequence classification problem. In this context, the goal is to classify the entire sequence (research paper or abstract) into predefined categories or classes. These classes can either represent different topics or keywords such as plant ecology, functional, ecology, species distribution models, or whether a paper is relevant to a particular database or study, By treating ecological paper identification as a sequence classification task, it becomes possible to leverage the power of modern natural language processing techniques to automatically categorize research papers, facilitating more efficient literature review and information retrieval processes within the field of ecology.

This approach was demonstrated by [24] to expand literature-based datasets. The authors train ML and DL models to classify whether literature is relevant to the Living Planet Database (LPD: http://livingplanetindex.org/data_portal) and the PREDICTS databases [54], with over 90% accuracy, significantly improving efficiency at which potentially relevant papers are discovered. The authors used methods like logistic regression and feed-forward neural networks to achieve this task. Within the thesis, we will use the LPD and PREDICTS datasets from that paper to demonstrate how large language models perform on the task using the same datasets. We will again evaluate the use of a logistic regression model on the same dataset to use as a baseline for the model performance, as our results may be different from the ones reported in the paper due to differences in data preprocessing, model training, among others. For each dataset, we conduct this task using two types of inputs. The first type utilizes only the paper titles to predict their relevance, while the second input incorporates the abstracts of the papers. This leads to a total of four dataset combinations on which we will assess the models. Given that this task involves sequence classification, the encoder architecture is the most appropriate. We assess the performance of four distinct encoder models: DistilBERT, EcoBERT, DeBERTaV3 and ELECTRA, each of which is trained on different corpora and training schemes.

3.1.2 Text summarization

Text summarization could potentially play a pivotal role in ecological literature surveys by significantly expediting and enhancing the process of literature synthesis. With the vast and ever-growing volume of ecological research publications, summarization techniques can allow researchers to distill the essential findings, methodologies, and insights from numerous papers into concise summaries [136, 14]. This not only saves time but also aids in comprehending and comparing research across various studies, facilitating the identification of trends, knowledge gaps, and emerging areas of interest within the field of ecology. Additionally, text summarization enables the extraction of critical ecological data and results, which can be invaluable for meta-analyses, evidence synthesis, and evidence-based decision-making in ecological research and conservation efforts. Summarized text can also be used as input in other NLP models in order to keep most relevant information, while dramatically decreasing the amount of text needed as input in the models. This in turn reduces the computational power necessary for such tasks, while keeping, or even increasing the model performance as texts which do not contain relevant information are truncated. This task, can be effectively conceptualized as a problem where the objective is to condense extensive ecological information, often found in research papers or environmental reports, into concise and coherent summaries. In this context, the input text, which may encompass full text or abstracts of scientific articles, is treated as the source document, and the goal is to generate a shorter, coherent, and informative summary that encapsulates the titles, key findings, insights, and implications present in the original text. By framing ecological information summarization as a text summarization task, it leverages advanced natural language processing techniques, such as sequence-to-sequence transformer-based models [14], to automatically extract and distill the most salient ecological knowledge, facilitating efficient data comprehension and knowledge dissemination within the ecological community.

In the thesis, we employ text summarization large language models to summarize the abstracts of papers into concise titles. To achieve this we utilize abstracts and titles from the PREDICTS and LPD datasets mentioned above. Furthermore, we evaluate the models using a dataset that includes abstracts and titles from bioarXiv. To gauge model performance, we compare it to a baseline where we extract the first three sentences of the abstracts as a summary. As text summarization is a sequence-to-sequence NLP task, where we input a text sequence into the model and get a text sequence as output, the encoder-decoder architecture is the optimal choice to explore. As a result, we focus on two encoder-decoder models: T5-FLAN and BART. However, there are numerous avenues for expanding and customizing this task in the future. For instance, we could extend the summarization to encompass the entire full text of papers or focus on summarizing specific sections of the text, extracting information relevant to particular aspects of the topic. Furthermore, decoder models can also be used for this task. They operate in an auto-regressive manner, generating a token (word) at a time, often resulting in a lower performance. Nonetheless, given the capabilities demonstrated by models like LLaMA[118], GALACTICA[115] and GPT-4 across a variety of tasks, investigating decoder models presents an intriguing prospect for future research.

3.2 Entity extraction

3.2.1 Named entity recognition

Named entity recognition (NER) in ecology holds immense significance for streamlining and optimizing information extraction from ecological literature[89, 77, 109]. The ecological domain witnesses a constant influx of research papers and reports, many of which contain valuable data regarding species, habitats, and ecosystem dynamics. Through the identification and classification of these ecological units inside the text using NER approaches, researchers can quickly access critical information without having to delve into lengthy papers. By automatically extracting species names, habitat descriptions, and other ecologically relevant entities, NER not only accelerates the literature review process but also enhances data comprehension and cross-study comparisons. All of this makes NER a valuable tool for data synthesis, evidence-based decision-making, and advancing our understanding of complex ecological systems.

Most taxonomic uses of NER have been limited towards identifying organisms in biomedical texts[40, 42]. However, seminal works have also focused on the extraction of taxonomic names from biodiversity legacy literature [102, 61]. Over the years, the developed taxonomic NER systems can be categorized into the following categories: rule-based, dictionary-based or based on ML. Most recently, a model called TaxoNERD[65] was created for this particular use of taxonomic NER and makes use of LLMs to achieve state-of-the-art results on four NER gold standard corpora. In the thesis, we expand on this work and utilize two of the datasets used in the TaxoNERD paper: SPECIES800 and COPIOUS, to test how NER task performance varies for different LLM architectures. NER can be framed as a token classification task, where for each word of the description we want to assign a class such as "species", "microorganism", "author" or no entity. Due to this, we focus on the encoder architecture across the following four models: DistilBERT, EcoBERT, DeBERTaV3, ELECTRA, in order to analyze how different pre-training approaches transfer to an ecological NER domain.

3.2.2 Family classification

By assigning each species to its respective taxonomic family through sequence classification, researchers can efficiently categorize and organize species data, facilitating the study of species distributions, biodiversity patterns, and ecological relationships. This automated classification not only saves time but also enhances data accuracy, especially when dealing with large datasets encompassing numerous species. Additionally, the predictions generated by sequence classification models can be integrated into broader ecological analyses, enabling the identification of taxonomic trends, ecosystem compositions, and species associations. While this task and related ones, such as species identification have been extensively researched in fields such as computer vision[125] and serve as the foundation for applications such as FloraIncognita[78] and Pl@ntNet[44], as far as we know, there has been no analogous research in the natural language domain. However, this approach has the potential to further aid these applications by providing a natural interface for the user to describe the species in cases where the vision model or user are uncertain about the species in question. By including model explainability in such tasks, the user would also receive detailed explanations of why the model predicted a certain family or species. In cases like this, these models can act as an addition to plant identification keys by highlighting what aspects to the species' characteristics make it unique. We take species' descriptions from two large online databases: Plants of the World Online and Wikipedia, which were used as input for the models. The descriptions were aligned to the corresponding species' family which was then as the label for the task. As a sequence classification task, we again use the four encoders models: DistilBERT, EcoBERT, DeBERTaV3 and ELECTRA, to evaluate how differences in their training approach result in the task performance.

3.3 Trait extraction

Functional trait extraction is crucial in ecological research as it provides valuable insights into how organisms interact with their environments and ecosystems. These traits can offer predictive power for understanding ecological processes, species distributions, and ecosystem functioning. NLP models offer immense potential in automating the extraction of functional traits from vast ecological datasets, enabling researchers to efficiently analyze and interpret trait data on a large scale, identify patterns, and advance our understanding of how species traits influence ecological dynamics and responses to environmental change.

NLP has previously been used for the extraction of traits [32, 86]. However, most of these approaches have stuck to simpler models such as dictionaries, term co-occurrences or bag of words models, which have major drawbacks in terms of complexity and predictive power when compared to large language models. In our previous work, we demonstrated that LLM's show outperform several of these models on the trait prediction task [30]. Furthermore, these extraction methods can be combined with methods from other fields such as CV, Remote sensing [9, 15] and Citizen Science data [134] to infer traits with a higher performance and confidence. For instance, by utilizing convolutional neural networks (CNN) on coupled images from iNaturalist and traits from the TRY database, [104] were able to accurately predict trait values. CNNs have also been used to measure functional traits of skeletal museum specimens [129]. The creation of methods for a cost-efficient and straightforward extraction of traits is instrumental to fix some of the geographic and socioeconomic biases outlined in the introduction. The utilization of text mining techniques from the AUS traits databases [33] was shown to increase the completeness of traits in Australia [79]. Specifically, through the use of these methods, the mean completeness of 53 traits in Australia grew from 9.9% to 23.8% and covered 89% of Australian vascular plants, raised from 46% of initial Australian vascular plants which had trait data in TRY. This highlights the potential of models like this to fill databases in a fast and smart way. Finally, correlations and phylogenetic relationships among species may help to further infer traits [105]. However, it is important to note that these methods have their limitations and are most effective when applied to taxonomically and geographically representative datasets.

Leveraging the power of NLP and complementary fields such as CV and Remote Sensing could potentially help us reach a necessary lower limit of known traits in order to be able to predict other traits through imputation methods [113]. We show how NLP models can be applied to extract two distinct types of traits: categorical and numerical. Furthermore, for categorical traits, we extend our investigation beyond the default scenario of extracting traits from English descriptions. We explore two other scenarios: the extraction of traits from Spanish and German descriptions, broadening the linguistic scope of our analysis. Finally, we try to tackle the challenge of trait extraction in data-deficient environments, characterized by a scarcity of labeled description data.

3.3.1 Categorical trait classification from English descriptions

We address the task of automated categorical functional trait extraction from text as a sequence classification problem. We begin with a species' textual description, which can be sourced from various places such as floras, academic publications, and plant databases. After performing standard NLP preprocessing, we feed the input into the transformers. The resulting output of the model is a predicted trait value along with a corresponding confidence score. We considered two databases which were created by combining species' descriptions from Plant of the World Online and English Wikipedia with traits from the GIFT database[131]. While this approach of merging data from two distinct databases has many merits, most importantly the speed at which we can acquire labeled data, it also has several downsides as the description data may not correctly represent the label [30]. We further focused our efforts on two categorical traits, using their definitions from the GIFT database. The first trait we considered was the plant's growth form which takes on three possible values: herb, shrub or tree. As mentioned in the introduction, growth form is the trait with the highest overall coverage in both the TRY and GIFT databases and therefore has a high chance to be specified in the species' descriptions. The second trait is the plant's life form, one of phanerophyte, chamaephyte, hemicryptophyte, cryptophyte or therophyte. This trait is much more complex for the model to predict as it has a very low coverage and is unlikely to be directly stated in the description. Therefore, the model has to predict the species' trait based on context clues, such as assigning the phanerophyte class to all trees. To benchmark the model's performance, we will evaluate it against a keyword search and logistic regression model. Furthermore, we will extend our previous work [30] where we utilized the DistilBERT architecture, by incorporating additional novel language models, including DeBERTav3 and ELECTRA, as well EcoBERT, a language model that has undergone further pre-training on ecological texts.

3.3.2 Categorical trait classification from non-English descriptions

Given that many trait descriptions are sourced from national floras or historical texts, a significant portion of this data may not be written in English but in other languages such as Spanish, Portuguese, or German. This represents a valuable source of data that cannot be effectively tapped into using English-based natural language processing alone, particularly for non-English speaking countries where most of the data-deficient species lie. To address this diversity in languages and mitigate the English-language bias [62], we explore the utilization of monolingual LLMs trained in languages other than English, as well as multilingual models trained on multiple languages at once. We employ different monolingual and multilingual large language models to predict categorical traits from two datasets: one containing Spanish Wikipedia descriptions, and another containing German Wikipedia descriptions. We first start with a English pre-trained DistilBERT model to access the extent to which knowledge learned by such a model would be able to transfer to other languages. Subsequently, we leverage a multilingual version of this model, ideally capable of performing effectively across multiple languages. Finally, we evaluate the performance of a monolingual Spanish and German pre-trained BERT. Similar to the English descriptions, we establish a baseline performance using a keyword search and logistic regression model for comparison. This multilingual approach aims to broaden the applicability of ecological trait extraction models and reduce language-based biases in ecological research.

3.3.3 Categorical trait classification in a data-deficient regime

In ecological research, dealing with limited data for certain traits presents a substantial and recurrent challenge. The scarcity of labeled training data can render the conventional supervised learning pipeline ineffective, as the model lacks sufficient training examples to discern the intricate relationships between input descriptions and trait classes. This issue necessitates the application of innovative techniques tailored to operate in data-deficient scenarios. To address this challenge, we will employ the approach of few-shot learning, specifically designed to thrive in conditions where data availability is restricted. By leveraging a small subset of the dataset, we aim to train models capable of accurately predicting categorical traits from species descriptions. To benchmark the efficacy of this few-shot learning approach, we will compare its performance against that of a standard sequence classification task, using an equivalent number of training examples. We start off with a training dataset size of 32 descriptions, and we increase this size four-fold within two iterations to a size of 128 and 512 samples. This comparative analysis will shed light on the utility and adaptability of few-shot learning in the realm of ecological text analysis, offering valuable insights into its potential for handling data scarcity in trait prediction tasks.

3.3.4 Numerical trait classification

Numerical traits are those described by a continuous numeric value. In the thesis, we focus on the three such traits that encapsulate important plant functional strategy information [137, 29]. Consequently, we adopted an extractive, or context-based, question answering (QA) model to constrain trait predictions within the context of the text. Extractive QA models function by posing a question (e.g., "What is the height of the plant?") and providing an answer along with a confidence score, all based on the information contained within a context paragraph, which, in our study, corresponds to the species' description. Additionally, the QA model returns both the numerical value and its associated unit of measurement, simplifying the process of converting between different units. To obtain the final predicted numerical value, we implemented post-processing steps, which involved filtering out answers lacking a numerical value or unit. Additionally, we set a confidence score threshold to exclude predictions with confidence scores below this threshold. In cases where the answer contained two numerical values, we extracted the second value, operating under the assumption that they represented a range (minimum-maximum). Finally, we ensured that the predicted value was transformed into the requisite unit of measurement for the trait in question.

4 Data

4.1 Literature review

To train and evaluate the topic modeling and text summarization models, we use the bioaRxiv, PREDICTS and LPD databases that contain information on paper titles and abstracts. The PREDICTS and LPD databases are taken from [24] and additionally contain information on whether they are relevant to their corresponding literature databases. The databases are acquired by taking positive examples: articles from the LPD and PREDICTS databases and negative examples: articles from the National Center for Biotechnology Information. These negative examples are pseudo-negative and are gathered using a process similar to gathering pseudo-negative examples for species distribution models. This means that it is not certain that these examples are actually irrelevant to the databases, but as they are taken from a dataset outside of them, the chance for this is relatively small, allowing for the model to learn this differentiation. This resulted in the LPD and PREDICTS databases used in the thesis which contain 5,633 and 5,536 abstracts and titles respectively. The bioaRxiv database was acquired from the bioaRxiv website (www.biorxiv.org) which contains scientific preprints in the fields of life sciences. We extract information on paper titles and abstracts using a web scraper between the years 2000 and 2023 in the subject area of Ecology, resulting in a total of 7,694 texts. The use of full-text manuscripts can be preferable to the use of abstracts in both the topic modelling and text summarization scenarios [132], however, they have a severely negative effect on the performance of the models, restricting the use of such models dependent on dataset size and computational resources. Furthermore, while not done here, another interesting task will be to summarize full text articles to their abstracts, which might be interesting to summarize larger text to a shorter summary.

4.2 English species' descriptions

For most of the above-mentioned tasks that deal with sequence classification and QA we require textual data and corresponding labels. To acquire the textual data we performed a web scrape of species' descriptions from two large online plant knowledge base, Plants of the World Online (POWO), which aggregates information from regional floras, and English Wikipedia (POWO), a community-written online encyclopedia. Plants of the World Online[3] is an online portal by Kew Royal Botanic Gardens that aims to digitize and share data on the world's flora. The data for each plant species includes images, taxonomy information, and textual descriptions of traits, identification, and distribution. These descriptions are organized hierarchically, with details on leaf morphology, plant habit, and reproductive information. To obtain these descriptions, we used the taxize R package [12, 11], resulting in 288,254 descriptions for 59,151 plants categorized into 251 distinct categories. Wikipedia[4] is a multilingual online encyclopedia that is written and maintained by a community of volunteers with material on a wide range of subjects, including plant species. To get species' descriptions from Wikipedia, we searched for English articles for around 200,000 species for

[3] http://www.plantsoftheworldonline.org/

[4] https://en.wikipedia.org/

which we have functional trait data in the GIFT database. We chose this direction because to train and evaluate the model, we need both textual descriptions and labels for each species. We used the Python Wikipedia-API[5], and the Requests and Beautiful Soup[6] web scraping libraries. This resulted in 194,994 descriptions for 55,631 species with description categories based on the sections in Wikipedia. To streamline the process and conserve computational resources, we amalgamated all descriptions per species from various sources to generate the intermediate datasets. Given the substantial number of models required for different tasks, we initiated the data preparation by removing duplicate entries and eliminating descriptions containing fewer than 10 words. Subsequently, to obtain the final POWO and WIKI datasets, we strategically sampled a subset comprising 5000 descriptions. This meticulous curation of data ensures the efficiency and effectiveness of subsequent model training and evaluation processes.

4.3 Spanish and German species' descriptions

The beauty of Wikipedia is that it also contains thousands of descriptions in a variety of languages, making it a valuable resource for the creation of datasets beyond the English language. We capitalize on this wealth of information to create datasets akin to the English Wikipedia dataset mentioned above. To do so, we searched the Spanish[7] and German[8] Wikipedia's for articles corresponding to 50,000 species with available functional trait data. This resulted in a collection of 24,531 Spanish descriptions for 4,196 species and 20,275 German descriptions for 2,419 species. As with the English datasets, all descriptions were aggregated per species per source to construct the final WIKI_ES and WIKI_DE datasets.

4.4 Taxonomic and functional data

To be able to use our textual descriptions in tasks related to taxonomic classification and functional trait extraction, the acquisition of labels for model training and evaluation is imperative. In address this need, we harnessed the resources provided by the Global Inventory of Floras and Traits (GIFT, [131, 27]) database, a comprehensive global repository comprising regional plant checklists, floras, and plant functional traits. GIFT boasts an extensive collection encompassing over 290,000 species and 109 distinct traits, rendering it an invaluable source for extracting and utilizing traits of interest as labels in our models. Using the GIFT R-package, we extracted trait values and employed them as labels within our model training pipeline. It is important to acknowledge, however, that the availability of labeled data exhibited variations among different traits within the database. This variation in label availability underscores the need for a thoughtful approach, as it can significantly influence the precision and effectiveness of models trained on such diverse and sometimes limited data.

[5] Wikipedia-API 0.5.4, Available at: https://pypi.org/project/Wikipedia-API

[6] beautifulsoup4 4.11.1. Available at: https://pypi.org/project/beautifulsoup4/

[7] https://es.wikipedia.org/

[8] https://de.wikipedia.org/

4.5 Named entity recognition

SPECIES, otherwise referred to as Species-800 or S800 [89], represents a pivotal addition to the realm of species name diversity, distinguishing itself from the LINNAEUS corpus. The genesis of S800 involved a deliberate effort to enrich the spectrum of species names by drawing from an eclectic array of sources. The foundation of S800 was laid by hand-selecting 100 MEDLINE abstracts from scholarly journals across eight distinct categories: bacteriology, botany, entomology, medicine, mycology, protistology, virology, and zoology. In a meticulous annotation process, taxonomic mentions encompassing Linnaean binomials, common names, strain designations, and author-defined acronyms were diligently identified and annotated. While the primary emphasis rested on annotating species mentions, other taxonomic ranks, such as kingdoms, orders, genera, and strains, were also thoughtfully considered. Remarkably, the S800 corpus boasts a nearly equivalent count of annotated species mentions as the LINNAEUS corpus[40], totaling 3,708 mentions. However, what sets S800 apart is its impressive repository of over three times the number of unique species names compared to its predecessor. The dataset used for the NER task consisted of 562 descriptions.

COPIOUS, as introduced in [88], emerges as a crucially specialized gold standard corpus (GSC), designed with a distinct focus on the extraction of species occurrences from the vast expanse of biodiversity literature. In contrast to its predecessors, LINNAEUS and S800, COPIOUS boasts a notably broader scope, encompassing a comprehensive array of entities. The corpus extends its reach to encompass taxonomic names, geographical locations, habitat descriptions, temporal expressions, and personal names. It draws its content from a substantial pool of 668 document pages meticulously culled from the Biodiversity Heritage Library, reflecting a diverse and comprehensive range of ecological literature. A team of experts embarked on the intricate task of manual annotation, meticulously tagging over 28,000 entities. Notably, 44% of these entities, totaling 12,227, pertain to taxa. The annotated taxon mentions span an extensive spectrum, encompassing species, genera, families, and all higher-order taxonomic ranks. A distinctive feature of COPIOUS annotation lies in its inclusivity, covering both contemporary and historical scientific names. For scientific names that incorporate authorship information, the corpus encapsulates two separate entities: one with authorship information and the other without. These entities overlap, sharing a common substring. It is noteworthy that COPIOUS was meticulously tailored for the specific purpose of extracting information related to Philippine biodiversity. Consequently, a portion of the common names included in the corpus represents English transcriptions of Filipino names, reflecting its regional focus. However, the authors underscore that the corpus exhibits a versatility that extends beyond its primary focus and can be effectively harnessed for various biodiversity-related applications. An additional delineation in COPIOUS is the exclusion of microorganism names, aligning with its distinct emphasis on highly endangered species and broader ecological considerations. The COPIOUS dataset used in the thesis had the same size as the previous dataset, also consisting of 562 descriptions.

4.6 Text preprocessing

All textual descriptions were preprocessed before being used as input in the models. The preprocessing pipeline consisted of the removal of artifacts and for the English texts, also the removal of accents from the text. The text was lowercased and split into tokens, which in our case may represent words, numbers, or punctuation.

5 Methods

5.1 Natural language processing in ecology - literature analysis

In order to quantify publication trends in the application of computer science methods in ecology, we used data from Web of Science (WOS)[9] searches (Fig. 6.1). We first used a search of 'ecology' OR 'biodiversity' in the Topic field combined with "NLP" OR "natural language processing" OR "text mining" to find papers that combine NLP with ecology, "CV" OR "computer vision" OR "image processing" OR "machine vision" to find papers that combine CV with ecology and "bioacoustic*" OR "acoustic*" OR "sound processing" to find papers that combine bioacoustics with ecology. Furthermore, to quantify how many papers, as well as what proportion of these papers, use foundation models, such as the transformer architecture, we performed a new search by combining the previous searches with 'transformer' OR 'LLM' OR 'large language model' OR 'foundation model' for the NLP domain, 'transformer' OR 'VIT' OR 'foundation model' for the CV domain and 'transformer' OR 'HuBERT' OR 'foundation model' for the bioacoustics domain. This search was done for papers after 2018, since the introduction of the transformer architecture was in December, 2017 [122]. While this isn't an optimal approach, since it can result in false positives: where publications mention these keywords without utilizing them and false negatives: where some papers that actually use these approaches aren't included, it should still give us a general insight on publication patterns in the fields.

5.2 Keyword search

For the categorical trait extraction tasks, we conducted keyword search, a widely adopted method in the automated extraction of traits. To do this, we formulated trait-specific dictionaries and employed a script driven by regular expressions to classify the descriptions. To simplify the dictionaries, the incorporated keywords comprised the class name (representing trait values) that could be encountered within the descriptions. The same was also done for the Spanish and German text descriptions by including the corresponding translated trait values in their dictionaries.

5.3 Logistic regression

We additionally train a logistic regression model for each sequence classification task in order to compare the large language models to a conventional machine learning approach. We opted for logistic regression as it is a parametric predictive classification model that

[9] https://webofscience.com/

Table 1: An example input text and output label or text for each of the explored natural language processing tasks.

Task	**Input**	**Output**
Topic Modeling	abundances of red fox and pine marten in relation to the composition of boreal forest landscapes	Relevant: TRUE
Summarization	conservation of naturally sympatric endangered species requires unique considerations while impacts of invasive species garner much attention interactions between endangered species must also be managed the endangered leon springs pupfish cyprinodon bovinus has suffered a population decline due to decreasing natural habitat as breeding habitat is lost c bovinus is also adversely affected by the sympatric endangered pecos gambusia	conservation and conflict between endangered desert fishes
Named Entity Recognition	Rapid Cold Hardening and Expression of Heat Shock Protein Genes in the B-Biotype Bemisia tabaci	Bemisia (B-Species) tabaci (I-Species)
Family Classification	A tall perennial ; Seeds with scattered groups of short stellate hairs ; Leaves on long hispid petioles; with conspicuous ear-like stipules ; Flowers orange-red. ; Covered with down and spreading hairs	Malvaceae
English Categorical Trait Classification	A straggling shrub (or ? woody climber) with branches up to 3 m. in length. ; Styles 2, 7–10 mm. long, usually glabrous.; Younger branches ± densely ferrugineous-pubescent, older ones glabrous, lenticellate. ; Calyx eglandular or with 1–2(–3) small circular glands , ± densely sericeous outside; lobes 2–2.5 (–3) mm. long, ovate to ovate-oblong, usually glabrous but shortly ciliate at margins. ; Petals yellow, obovate, ± 12 mm. long, entire or shortly lacerate, clawed. ;	Growth Form: Shrub
Spanish Categorical Trait Classification	El abeto es un árbol casi piramidal fuerte, que alcanza alturas de 25 a 30 metros y un diámetro de 75 a 210 centímetros. La corteza es inicialmente suave y en las ramas jóvenes es de color amarillo-rojizo a gris-verde oliva. Las hojas son estrechas como agujas de color verde brillante, . El periodo de floración es en mayo, las semillas maduran en septiembre-octubre Los conos cilíndricos . de color rojizo-marrón, tienen una longitud de 15 a 30 cm y un . diámetro de 4 a 6 cm. Las semillas aladas son de color rojizo-marrón, y de aproximadamente 5 cm de largo	Growth Form: Tree
Numerical Trait Classification	Climber, sometimes shrub-like, up to 3.5 m. high, with the basal part c. 1 × 0.4 m. above ground, thickened and ± fleshy, emitting numerous stems up to 3 m. long .	Plant Height Max: 3.5 m

enjoys widespread popularity in ecological research. This model generates probabilistic predictions for each class under the assumption of a linear relationship between the independent variables and the target variable. However, a critical prerequisite for this model is numeric input. Thus, we initiated the process by transforming the textual input into a numeric vector-space by calculating a text embedding. To accomplish this transformation, we employed the widely recognized bag of words (BOW) technique, commonly known as one-hot encoding. The BOW method operates by initially identifying the most frequently occurring words (in our case, we selected the top 1000) within the entire textual corpus, thereby establishing the model's vocabulary. Subsequently, each description underwent transformation into a vector of size 1000, where each value corresponds to the frequency of the associated term within the description. This BOW representation vector was then integrated as the predictor within the logistic regression model, with the trait value serving as the desired outcome.

5.4 Language models

We harnessed a diverse array of language models that have undergone distinct pre-training methodologies and have been trained on disparate corpora. This strategic selection was driven by the specific requirements of each task we undertook. Depending on the nature of the task at hand, we leveraged both encoder and encoder-decoder models, tailoring our choice to optimize performance and efficiency in accordance with the unique demands posed by each ecological analysis we conducted.

- **DistilBERT**: The DistilBERT model[101], is a encoder-based transformer model that is trained in a self-supervised manner using knowledge distillation from the BERT model [28]. This approach allows DistilBERT to achieve similar results to BERT with significantly fewer parameters, only 66 million compared to BERT's 340 million. The model is pre-trained using masked language modeling and trained with distillation loss and cosine embedding loss, resulting in prediction probabilities and hidden states that closely match BERT's. The pre-training texts are the same as BERT, which includes the Toronto Book Corpus and English Wikipedia Corpus, providing general knowledge vocabulary and language model. To use DistilBERT in a categorical trait pipeline, a sequence classification head is attached and fine-tuned using species descriptions and trait data. Therefore, in summary, only fine-tuning is necessary to obtain the final model.

- **EcoBERT**: The pre-trained EcoBERT model builds on the general knowledge of DistilBERT by using it as its tokenizer and language model. However, for this model we further pre-trained the language model on an ecological corpus. The specific checkpoint of EcoBERT used was specifically pre-trained using a masked language modelling objective on titles, abstracts and whenever available full texts of articles in the fields of Ecology acquired from bioarXiv and JSTOR. This approach should use the semantic and syntactic general knowledge text representation from BERT and further strengthen these relations for ecological knowledge. Therefore, in short, in this model we apply both pre-training and fine-tuning to obtain our final model.

- **DeBERTaV3**: DeBERTaV3 [50], an evolution of the DeBERTa (Decoding-enhanced BERT with disentangled attention) model [51], represents a significant stride forward in the field of natural language processing. Building upon the foundations laid by its predecessors, DeBERTaV3 introduces several crucial enhancements, such as a pioneering training technique known as "shortcut attention," designed to expedite training convergence. Moreover, this iteration incorporates adaptive activation functions, further refining the model's efficiency. A distinctive hallmark of DeBERTaV3 lies in its exceptional capability to adeptly capture and model intricate linguistic dependencies, endowing it with substantial prowess for an extensive array of NLP tasks. These tasks encompass text classification, sentiment analysis, question answering, and more, all while exhibiting noteworthy improvements in both performance and computational efficiency. DeBERTaV3 exemplifies the continual evolution of transformer-based architectures, furnishing state-of-the-art capabilities that significantly enhance natural language understanding across diverse applications.

- **ELECTRA**: The ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately) [21] model represents a groundbreaking approach to pre-training in natural language processing. Instead of the traditional masked language modeling objective, ELECTRA employs a novel "discriminator" task where a subset of tokens are replaced with plausible alternatives, and the model learns to differentiate between genuine tokens and these replacements. This approach not only significantly accelerates training but also enables the model to capture subtle semantic relationships and contextual nuances. ELECTRA has demonstrated remarkable performance on various downstream NLP tasks, showcasing its efficiency and effectiveness in comparison to conventional pre-training methods. This innovative model has spurred advancements in transformer-based architectures and furthered our understanding of effective pre-training techniques in NLP.

- **Multilingual DistilBERT**: The Multilingual DistilBERT model is designed for multilingual natural language understanding tasks. Instead of pre-training the language model on the Book and English Wikipedia Corpus, this model is trained on the concatenation of Wikipedia in 104 different languages, including Spanish and German which are of our interest. By pre-training on a diverse range of languages, Multilingual DistilBERT can capture cross-lingual relationships and transfer knowledge effectively across multiple languages, enabling it to perform tasks like text classification, language modeling, and sentiment analysis in a wide array of linguistic contexts. This model has proven valuable for researchers and practitioners seeking multilingual solutions without sacrificing computational efficiency.

- **Spanish BERT**: The Spanish BERT model BETO[10] is a specialized language model designed to excel in Spanish natural language understanding tasks. It is a variant of the original BERT model, fine-tuned specifically for the Spanish language by training with the whole word masking technique. BETO captures the intricate nuances of Spanish text, enabling it to perform tasks like sentiment analysis, text classification, and language modeling with remarkable accuracy and context-awareness. It has become an invaluable tool for researchers, developers, and businesses operating in Spanish-speaking regions, offering state-of-the-art performance in a variety of NLP applications while enhancing our understanding of the Spanish language's complex linguistic structures.

- **German BERT**: The German BERT model [13] is a dedicated language model by the MDZ Digital Library team at the Bavarian State Library tailored to excel in German natural language understanding tasks. Again derived from the original BERT, this specialized variant has been fine-tuned specifically for the German language, enabling it to capture the intricacies and nuances of German text. The source data for the model consists of a recent Wikipedia dump, EU Bookshop corpus, Open Subtitles, CommonCrawl, ParaCrawl and News Crawl. This results in a dataset with a size of 16GB and 2,350,234,427 tokens. German BERT empowers a wide range of applications, including text classification, sentiment analysis, and language modeling, by providing a contextually aware understanding of the German language. This model has proven invaluable for researchers, businesses, and developers operating in German-speaking regions, offering high-performance natural language processing capabilities and enhancing our comprehension of the complexities inherent to the German language.

- **FLAN-T5**: The FLAN-T5 (Few-shot Language Adaptation with New Tokens for T5) model [20] is an innovative extension of the T5 (Text-to-Text Transfer Transformer) architecture, specifically designed to tackle the challenges of few-shot learning and adaptability to new languages. FLAN-T5 leverages a novel approach that allows it to integrate new tokens into its vocabulary, enhancing its capability to adapt to and generate text in multiple languages with limited training examples. This model not only excels in a variety of text generation tasks but also exhibits remarkable versatility in adapting to new languages and tasks with minimal effort. FLAN-T5 represents a significant advancement in multilingual and few-shot learning, providing a powerful tool for tasks such as translation, summarization, and question answering across diverse linguistic contexts.

- **BART**: The BART (Bidirectional and Auto-Regressive Transformers) model [68] is a versatile and powerful transformer-based architecture designed for various natural language processing tasks. What sets BART apart is its unique dual nature, combining bidirectional and auto-regressive capabilities in a single model. BART's bidirectional encoder captures contextual information from both directions, while the auto-regressive decoder generates text, making it well-suited for tasks such as text summarization, language generation, and text completion. Its pre-training process involves denoising text by randomly masking and reconstructing segments, leading to remarkable results in text generation and understanding tasks. BART has demonstrated its prowess in a wide range of applications, from document summarization to text translation, establishing itself as a pivotal model in the transformer-based NLP landscape.

- **MPNet**: The MPNet (Masked and Permuted Pre-training) architecture [110] represents a pioneering addition to the realm of transformer-based models. What sets

MPNet apart is its unique approach to pre-training named masked and permuted language modeling, that involves masking and permuting text segments to instill a robust understanding of language. This approach equips MPNet with the adaptability needed to perform effectively with limited labeled data, making it well-suited for tasks such as few-shot classification, paraphrasing, and question-answering in resource-constrained settings. For the few-shot learning task, we use a checkpoint of this model designed for sentence embeddings and paraphrase identification tasks, and apply contrastive learning using a cosine similarity loss.

5.5 Evaluation

For our ecological text summarization task, we employ ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [69] as a fundamental metric for evaluating the quality of our generated summaries. ROUGE is particularly well-suited for this task because it measures the similarity between the model-generated summaries and reference summaries, assessing how well the generated text captures the essential content and linguistic nuances of the original ecological documents. Specifically, we utilize four primary ROUGE metrics: ROUGE-1, which measures the overlap of unigrams (single words) between the generated summaries and reference summaries; ROUGE-2, which extends this evaluation to bigrams (two-word sequences); ROUGE-L, which evaluates the longest common subsequence, and ROUGE-L-SUM, which combines the ROUGE-L metric with a length-based penalty, favoring summaries that are concise yet informative. By utilizing ROUGE, we can quantitatively evaluate the summarization model's precision, recall and particularly for our case, its f1-score, allowing us to gauge its ability to extract and convey the most critical ecological insights and findings.

In tasks related to sequence classification, our primary evaluation metric of choice was the f1-score, which offers a more comprehensive assessment than accuracy, particularly in cases involving imbalanced datasets. Imbalanced datasets, where descriptions are unevenly distributed across various trait values, can lead to artificially high accuracy scores. To address this, we focused on precision, recall, and the f1-score, all of which are calculated independently for each class. Precision represents the proportion of true positive predictions out of all positive predictions made by the model, quantifying the model's ability to correctly identify relevant trait values. Recall, on the other hand, measures the proportion of true positive predictions out of all actual positive instances, assessing the model's capability to capture all relevant trait values without missing any. The f1-score combines these two metrics into a single value, emphasizing the model's capacity to strike a balance between precision and recall. By calculating precision, recall, and the f1-score for each class separately and then averaging them using the macro-average approach, we obtained a comprehensive understanding of the model's performance across diverse trait categories, effectively considering both precision and recall to provide a well-rounded assessment.

To assess the performance of the numerical trait models, we initially applied a logarithmic transformation to the trait values due to the highly skewed distribution of the data. To ensure comparability across different datasets and models, we further normalized the data to a standardized range between 0 and 1. The primary metric employed for evaluation was the normalized mean absolute error (NMAE). In addition to NMAE, we investigated another crucial metric that we define as "coverage." Coverage measures the proportion of predictions that surpass a predefined threshold and successfully yield a trait value along with its corresponding unit of measurement. This metric provides insights into the model's ability to generate meaningful predictions within a certain confidence range.

In our paper, we rely on Precision, Recall, and particularly the F1 score as fundamental evaluation metrics to assess the performance of our Named Entity Recognition (NER) model in the context of ecological data extraction. The F1 score is of particular significance as it offers a comprehensive evaluation of our NER model's performance, considering the harmonious balance between precision and recall. This metric serves as an overarching indicator of how effectively our model can accurately identify ecological entities, such as species names or habitat descriptions, within the text. High F1 scores reflect the model's proficiency in minimizing false positives while simultaneously ensuring the capture of a wide range of ecologically relevant entities. By utilizing the F1 score alongside other metrics, our paper rigorously evaluates the efficacy of our NER model, emphasizing the importance of achieving a balanced trade-off between precision and recall to enhance the overall accuracy and comprehensiveness of ecological data extraction within ecological research and literature analysis.

All datasets were split into a training set, consisting of 75% of the textual data, and a test set of the remaining 25%, unless they already had a predefined split such as in the NER datasets.

5.6 Implementation

The codebase was written in Python v. 3.9.13 and either directly or indirectly relied heavily on the NumPy [49] and pandas [82] libraries for the organization and processing of data. Text analysis and preprocessing was done using the NLTK python library [74]. We implemented the BOW and logistic regression model using the scikit-learn ML library v. 1.1.3 [91]. We trained and evaluated the large language models using the Huggingface's transformers library v. 4.28.0 [135], sentence transformers library [100] and the simple transformers Python library[10]. To train and evaluate the few-shot models we used the SetFit library [120]. The models were trained using Google Colab and Kaggle on the freely available T4 GPUs. While there were some differences in model training, the models were generally trained for 3 epochs with a batch size of 8 to 16 and a maximum sequence length of 512 tokens. The default learning rate was set to $2e^{-5}$ and a weight decay of 0.01 was applied. The visualizations were done using the Python Matplotlib [55] and Seaborn [128] libraries. The entire codebase for the thesis is open-source and available on GitHub[11].

[10] Simple Transformers 0.60.0 Available at: https://github.com/ThilinaRajapakse/simpletransformers

[11] https://github.com/ViktorDomazetoski/Ecological-LLMs

6 Results

6.1 Natural language processing in ecology - literature analysis

Out of all of the fields, bioacoustics had the most articles, with 3377 papers in total, 190 of which also discuss foundation models. There were 1831 papers that addressed computer vision and 92 of them utilized foundation models. NLP fares the poorest, with a total of only 157 papers, even less than the amount of bioacoustics paper which utilize foundation models. Out of these papers, only 15 use large language models.

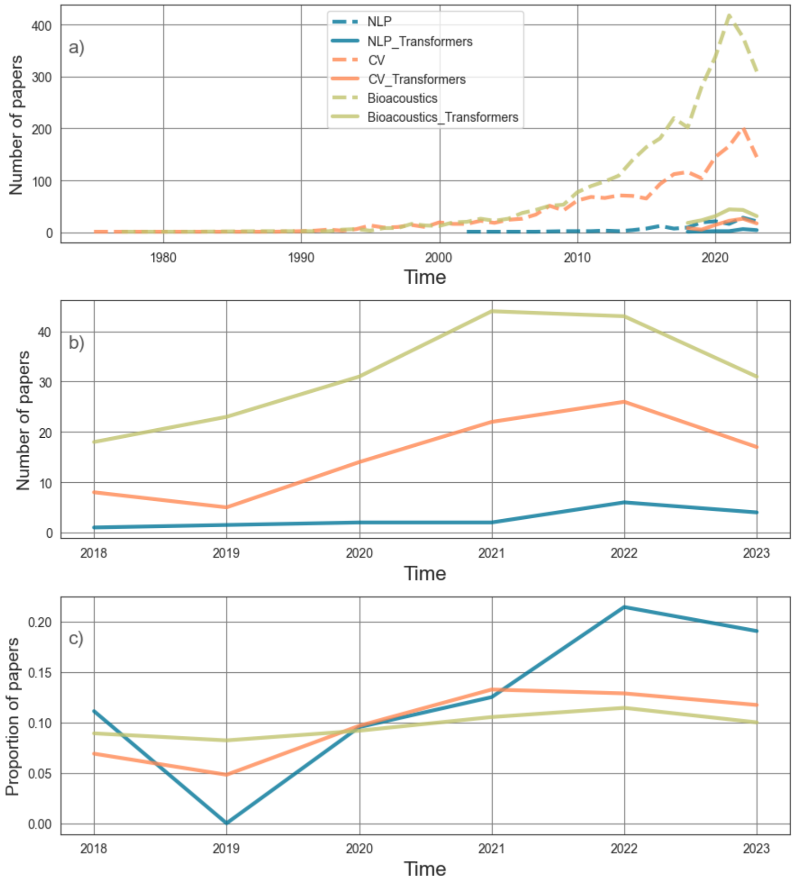

Figure 3: Publication trends showing the number of papers in the fields of natural language processing (NLP) (blue), computer vision (CV) (orange) and bioacoustics (yellow), without the utilization of the transformer architecture (a, dashed line) and with the utilization of the transformer architecture since 2018 (a,b, solid line). The proportion of papers that utilize the transformer architecture in their corresponding fields since 2018 is also shown (c). Data was taken from Web of Science (WOS) searches. Searches were conducted on the 10th of September, 2023.

6.2 Literature review

All of the models performed well on the topic modelling task, consistently achieving accuracy rates surpassing 93% (Fig. 4). When examining the performance difference between using only the paper titles and incorporating the entire abstract, it was observed that the models exhibited slightly lower performance when relying solely on titles. Specifically, when utilizing titles, the average F1-score among all LLMs reached 84.4%, while this score grew to 91.8% when utilizing the abstract. It is worth noting that the use of full text articles may unlock the potential for even higher performances. The PREDICTS dataset was overall less difficult, resulting in a 85.7% and 92.2% F1-score on the text and abstract data, whereas the LPD dataset yielded values of 83.1% and 91.1%. The best performing models were EcoBERT and DistilBERT, which achieved equal performance: 82.9% on LPI titles, 91.6% on LPI abstracts, 86.7% on PREDICTS titles and 93.3% on PREDICTS abstracts. However, the differences between the different large language models were minute, and certain models like ELECTRA achieved a slightly higher F1-score on specific tasks such as LPI titles exhibiting a F1-score of 83.5%, attributable to an increased recall and reduced precision compared to the other models. Across all datasets, LLMs consistently outperformed the logistic regression model. This difference was particularly pronounced on the LPD dataset, where there was an approximate 7% increase in F1-score when exclusively utilizing paper titles, along with a 6% increase when utilizing abstracts to predict whether the paper is relevant to the corresponding database. In the case of the PREDICTS database, these improvements amounted to 2% and 5%, respectively.

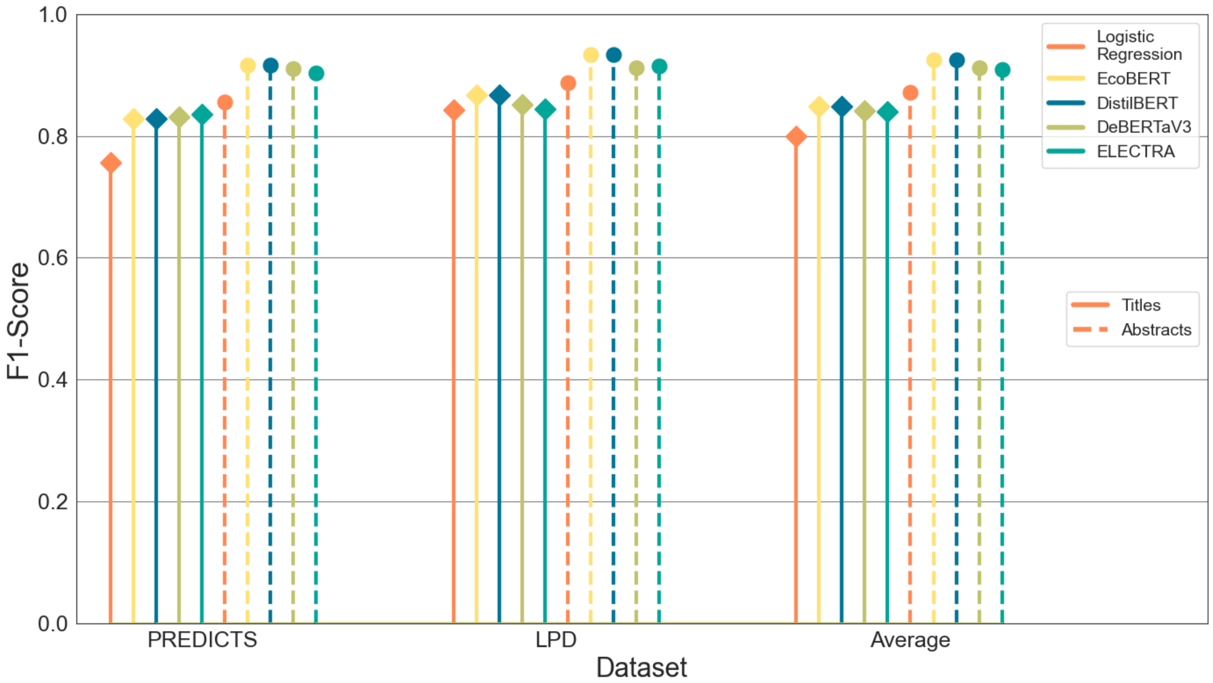

Figure 4: Topic modeling model comparison on the PREDICTS and LPD datasets. The f1-score is shown for the logistic regression (orange), EcoBERT (yellow), DistilBERT (blue), DeBERTaV3 (olive) and ELECTRA (cyan) models for the title (solid line diamond) and abstract (dashed line circle) texts.

The 3-sentence summary resulted in a baseline performance of ROUGE-L-SUM of 15.7%, 8.7% and 8.7% on the bioaRxiv, LPD and PREDICTS datasets, respectively. When considering the ROUGE-1 score, which measures the f1-score based on the overlap of single words (unigrams) between the generated and reference summaries, the average was equal to 13.65% across all datasets, while the average ROUGE-2 f1-score was notably lower at 5.6%. The large language models improved these scores significantly, with performance gains exceeding 200% to 300%, showcasing their capability in the task. The FLAN-T5 model achieved an average ROUGE-1 score of 39.6%, ROUGE-2 score of 17.8% and ROUGE-L-SUM score of 33.7%. BART was the best performing model, increasing these scores to 41.5%, 19.9% and 35.8%. While these scores indicate a moderate degree of content overlap between the generated and reference summaries, their interpretation is task-dependent and therefore further benchmarks and human evaluations need to be considered to further interpret the results in an ecological context. Across the datasets, bioaRxiv posed the greatest summarization challenge, with the BART model achieving a 32.8% ROUGE-L-SUM score, compared to the 37.5% and 37% scores on the LPD and PREDICTS datasets. This discrepancy is intriguing, given that the 3-sentence baseline on the same dataset was almost double the score of the two other datasets, prompting further investigation into the underlying factors contributing to this variation.

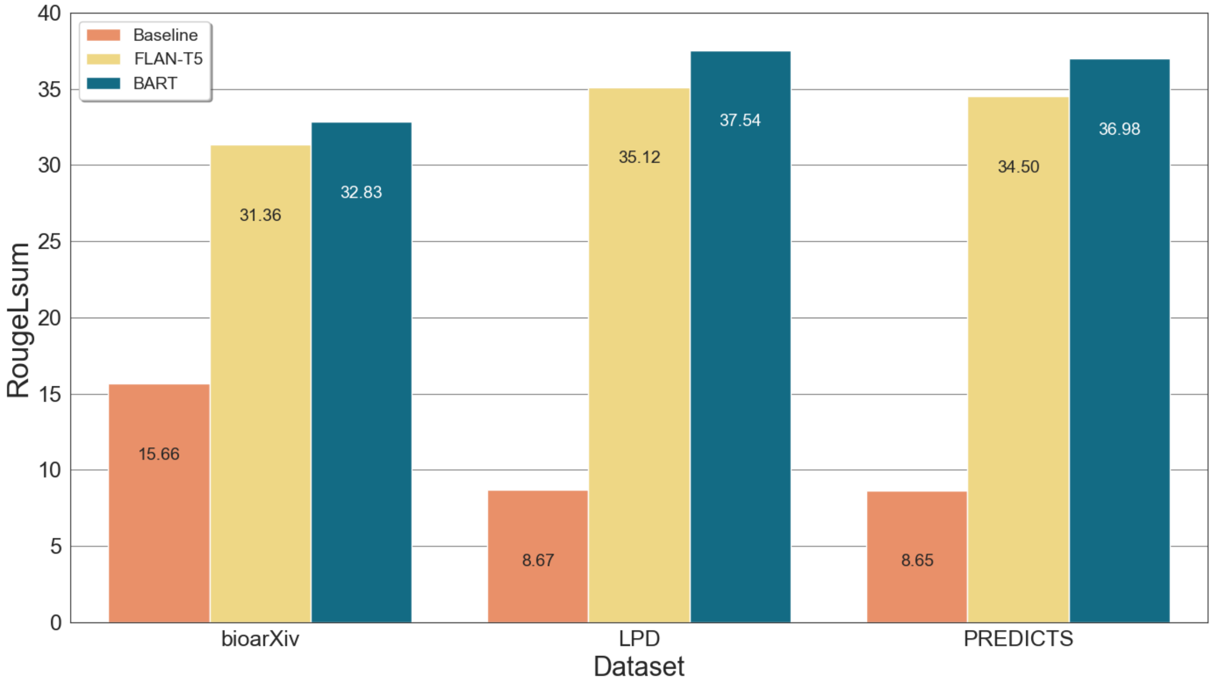

Figure 5: Text summarization results on the bioarXiv, PREDICTS and LPD datasets. The ROUGE f1-score is shown for the baseline three-sentence summary (orange), FLAN-T5 (yellow) and BART (blue).

6.3 Entity extraction

On the NER task, DeBERTaV3 emerged as the best model, achieving an F1-score of 80.9% on the COPIOUS dataset and 67.3% on the SPECIES dataset. ELECTRA also showcased commendable results, achieving F1-scores of 73.6% and 61.4%. These findings underscore the substantial performance enhancements realized in newer transformer architectures for NER tasks, signifying a notable leap from the DistilBERT model, which recorded F1-scores of 60.5% on the COPIOUS dataset and 53.3% on the SPECIES dataset. The EcoBERT model slightly benefited from its ecological pre-training, resulting in score of 61.4% on the COPIOUS dataset, albeit with a slightly diminished score of 52.8% on the SPECIES dataset (Fig. 6). It is worth noting that the results we have presented are subject to various factors, such as model evaluation methodologies. However, it is significant to highlight that the exact match F1-scores we achieved using the DeBERTa and ELECTRA models surpass those reported for the TaxoNERD BioBERT model [65], which were equal to 75.2% and 54.8% for the COPIOUS and SPECIES datasets, respectively, signifying a notable advantage of these language models.

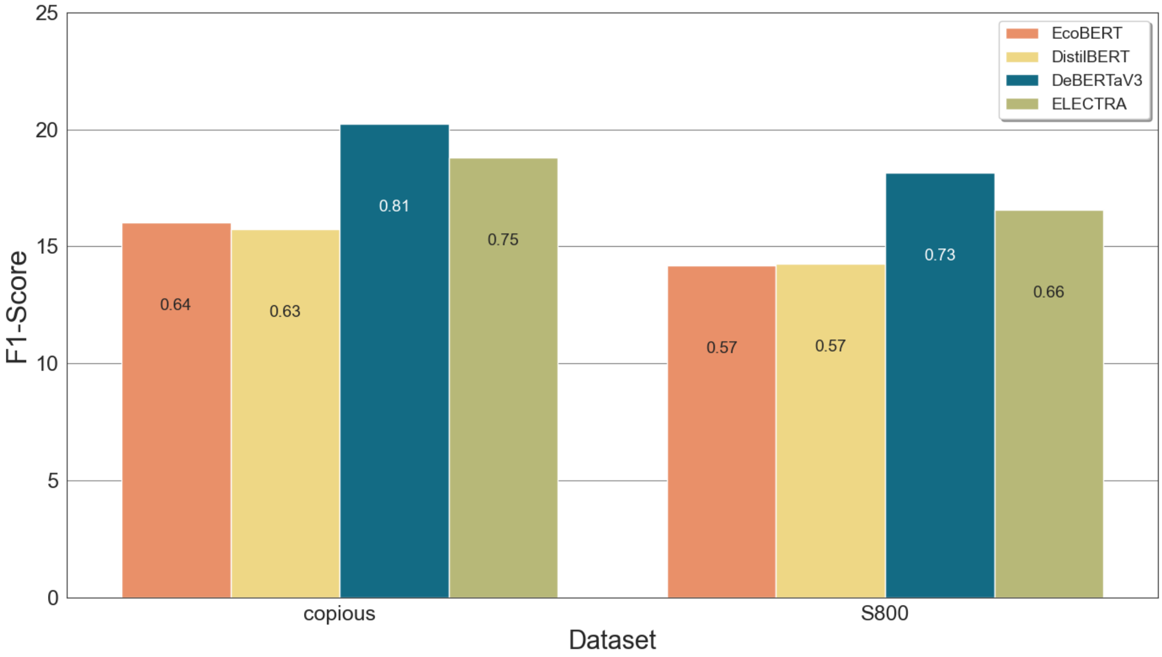

Figure 6: Named entity recognition model performance on the COPIOUS and S800 datasets. The f1-score is shown for the four fine-tuned language models: EcoBERT (orange), DistilBERT (yellow), DeBERTaV3 (blue) and ELECTRA (olive).

In the domain of family classification task, all models showcased exceptional performance, achieving F1-scores surpassing 95% (Fig. 7). DistilBERT secured the highest F1-score on the POWO dataset, registering a score of 96.8%, while ELECTRA had the highest score of 97% on the WIKI dataset. Although the margin of difference between the models was relatively small, the large language models exhibited a slight improvement in performance compared to the logistic regression model, with gains of 0.7% and 0.9% on the two datasets, respectively. This outcome underscores the model's competence in distinguishing plant families based solely on characteristics extracted from unstructured text. Looking ahead, this capability could potentially be extended to genus or species levels, serving as a valuable complement to plant identification keys for field species identification.

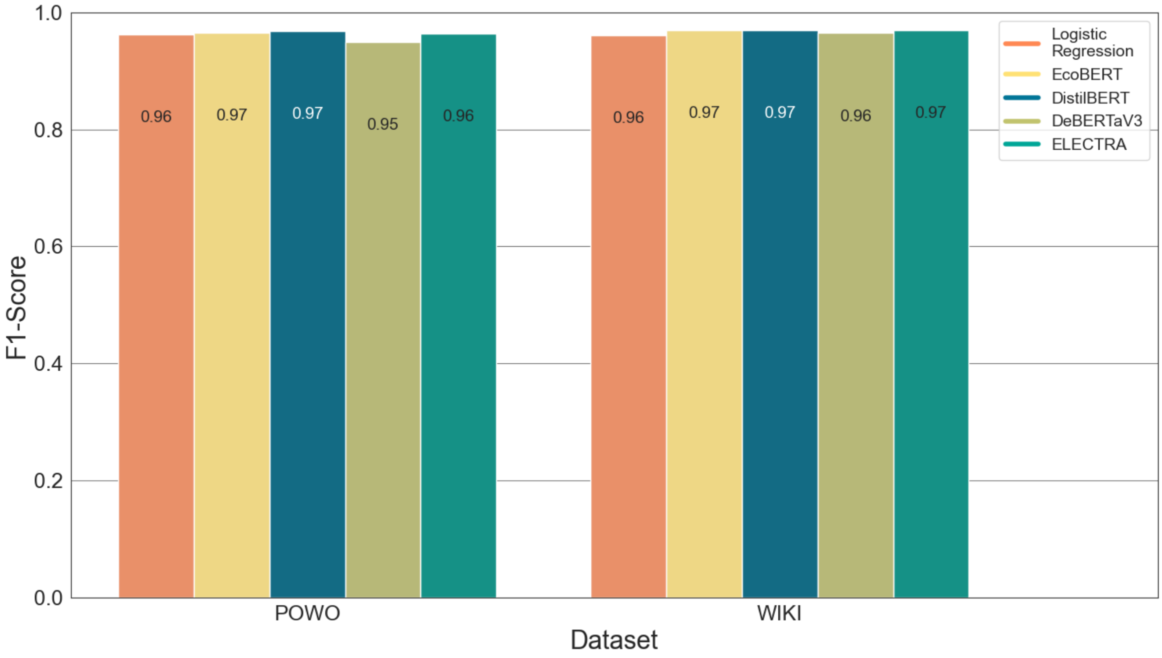

Figure 7: Results on the family classification task on the POWO and WIKI datasets. The f1-score is shown for the logistic regression (orange), EcoBERT (yellow), DistilBERT (blue), DeBERTaV3 (olive) and ELECTRA (cyan) models.

6.4 Trait extraction

The transformer models demonstrated superior performance compared to both the standard keyword search and logistic regression models. Specifically, on the POWO dataset, the LLMs achieved an average F1-score of 85.7% for the growth form trait and 83.9% for the life form trait. In contrast, the keyword search model yielded scores of 60.5% for growth form and 0% for life form, meaning that a substantial portion of the descriptions did not contain explicit information about the growth form, despite it being the most commonly reported trait. Moreover, none of the descriptions provided information about the life form, emphasizing the limitations of this approach. The logistic regression model also underperformed, with F1-scores of 79.3% and 81.9% on the two traits. When evaluating individual LLMs, the EcoBERT model emerged as the top performer, achieving F1-scores of 86.4% for growth form and 86.9% F1-score for life form. This underscores the importance of domain adaptation and the value of a model infused with ecological knowledge. Surprisingly, the DeBERTaV3 and ELECTRA models performed even worse than the DistilBERT model, despite their novel training approaches.

A similar trend emerges when analyzing the WIKI dataset. The keyword search model yielded F1-scores of 59.2% for growth form and a mere 0.6% for life form, reaffirming the limitations of this commonly used approach when dealing with such descriptions. LLMs again outperform the logistic regression model, demonstrating an approximate improvement of 10% on the growth form and 12% for the life form traits. In the case of the growth form trait, DistilBERT performed the best, achieving a remarkable score of 90.1%. However, it is worth noting that all models performed quite well on this trait, with scores ranging between 88% and 89.3%. On the other hand, for the life form trait, the EcoBERT model stood out with an impressive score of 74.2%, surpassing all other models, whose scores ranged between 64.2% for the DeBERTa model to 70.3% for the DistilBERT model. This further underscores the significance of pre-training language models on ecological corpora, as it can substantially enhance performance across various ecological tasks.

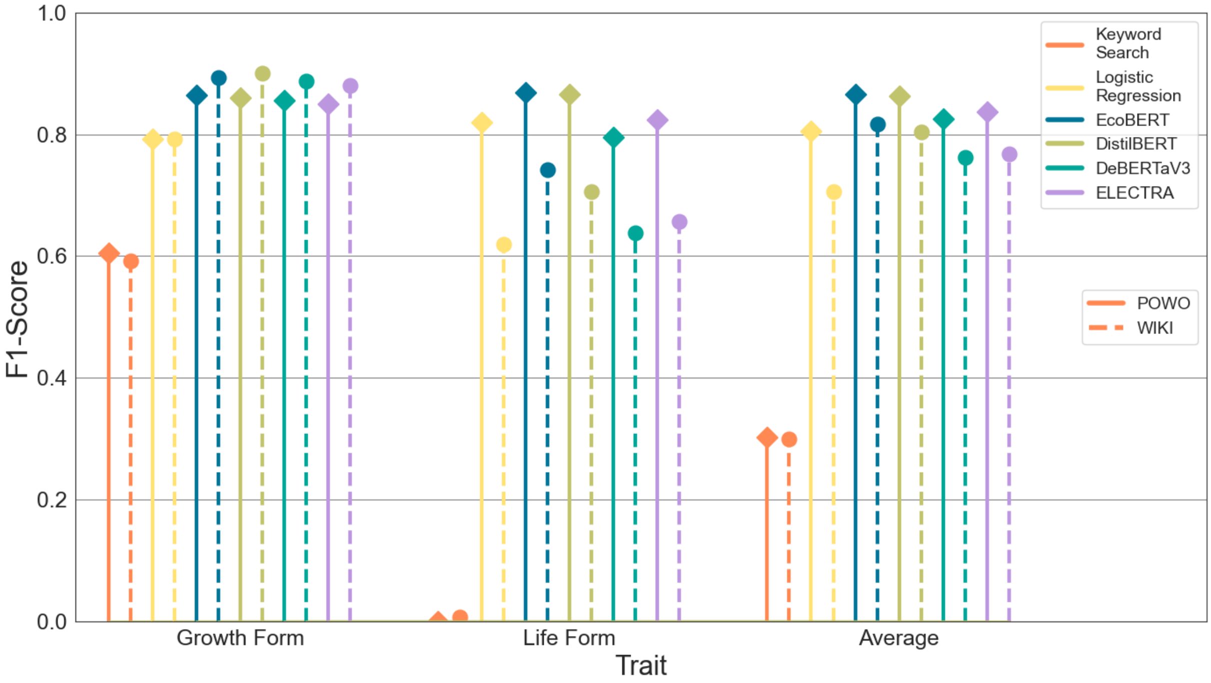

Figure 8: Model comparison on categorical trait classification of English descriptions. The f1-score is shown for the growth form and life form traits for the following models: keyword search (orange) logistic regression (yellow), EcoBERT (blue), DistilBERT (olive), DeBERTaV3 (cyan) and ELECTRA (violet) models for the POWO (solid line diamond) and WIKI (dashed line circle) datasets.

On the non-English Wikipedia datasets, as anticipated, the standard DistilBERT model exhibited notably poorer performance due to its training on English corpora. The F1-scores for the Spanish dataset were 81.4% for growth form and 59.4% for life form. Switching to the multilingual DistilBERT model, the growth form score dipped slightly to 81.3%, however the life form score unexpectedly dropped to 37.6%. In contrast, the BETO model excelled, achieving scores of 81.8% and 67.5% on the corresponding traits. A parallel trend emerged on the German dataset, where the base DistilBERT model yielded scores of 82.3% and 52.2%. While the multilingual model improved the growth form F1-score to 85.3%, it once again resulted in a decrease for the life form F1-score, which dropped to 47.3%. The German BERT model performed the best in this context, achieving scores of 87.4% for growth form and 82.2% for the life form trait.

In summary, the BETO and German BERT models demonstrated competitive performance on their respective datasets, akin to the large language models on the English Wikipedia dataset, This highlights their potential for knowledge discovery across various languages. Furthermore, these models outperformed the two benchmark models significantly. The keyword search model had a F1-score of 28.3% on the Spanish and 48% on the German datasets for the growth form trait, and a meager 0% and 8.9% F1-score on the life form trait. Comparatively, the logistic regression model yielded F1-scores of 75.9% and 61.5% F1-score on the Spanish dataset, and 79.4% and 69.5% on the German dataset.

In our few-shot experimentation, we deliberately reduced the amount of training data while keeping the entire test set intact. This approach aimed to assess how well the models

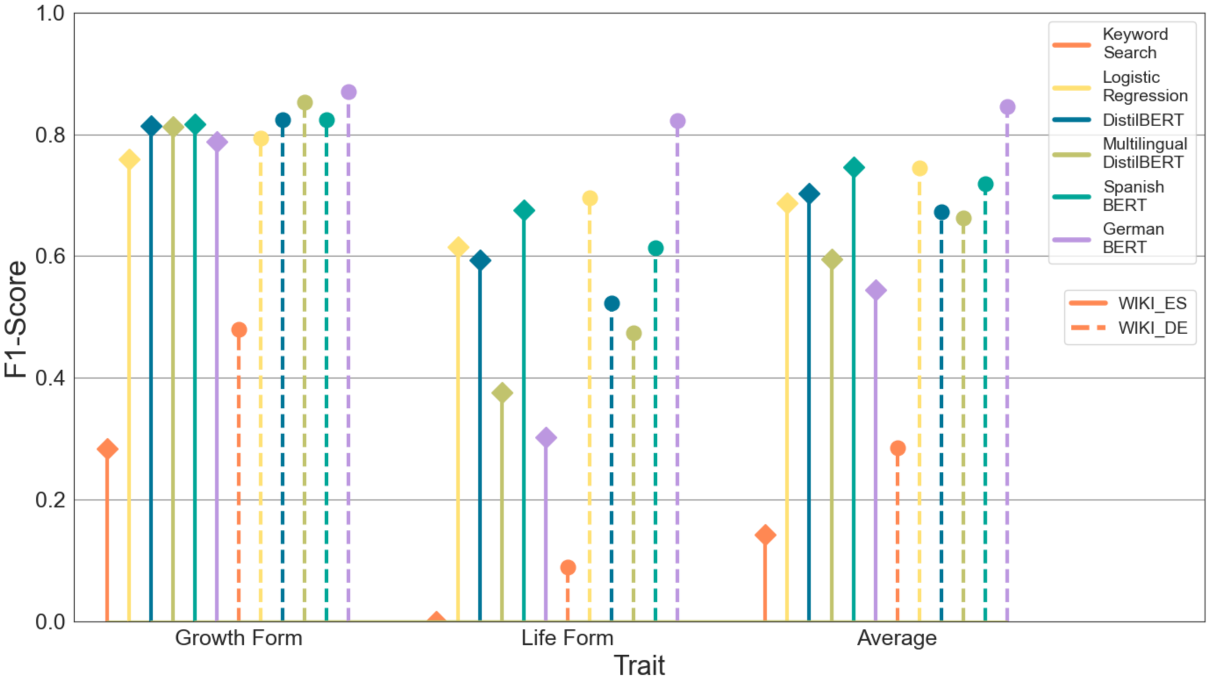

Figure 9: Model comparison on categorical trait classification of the Spanish (solid line diamond) and German (dashed line circle) descriptions. The f1-score is shown for growth form (orange), life form (yellow) and average (blue). We used the following models: keyword search, logistic regression, DistilBERT, Multilingual DistilBERT, Spanish BERT and German BERT.

could perform with limited training examples. When DistilBERT was trained with only 32 descriptions, it achieved a F1-score of merely 27.7% for the growth form trait on the POWO dataset, and an even lower F1-score of 22.9% on the WIKI dataset. Given the three possible classes in this scenario (herb, shrub or tree), this result was even lower than randomly assigning a class for each prediction, where we would expect a score near 33%. Upon increasing the training data to 128 descriptions, the performance on the POWO dataset remained stagnant, while there was a slight improvement on the WIKI dataset, reaching 23%. Subsequently, when the dataset size was again expanded by a factor of four, we observed significant performance gains, resulting in F1-scores of 79.7% and 83.9%, which approached the F1-scores of 86.1% and 90% achieved when using the entire dataset. Nevertheless, it is crucial to note that for most traits in the GIFT database, there are fewer than 512 training samples, highlighting the limitations of the standard sequence classification approach in such a data deficient scheme. In contrast, when we employed a few-shot training approach, the model exhibited slightly higher F1-scores on the 32 descriptions, reaching 30.7% on the POWO dataset. The improvement was more pronounced on the WIKI dataset, with a score of 51.9%. As we increased the size of the dataset to 128, these scores continued to rise to 79.5% and 82.2%, which were already competitive with those achieved with the 512-sample subset using the DistilBERT model. For the final dataset size of 512, the MPNet few-shot model yielded F1-scores of 83.5% on the POWO dataset and 83% on the WIKI dataset, demonstrating the effectiveness of this approach (Fig. 10).

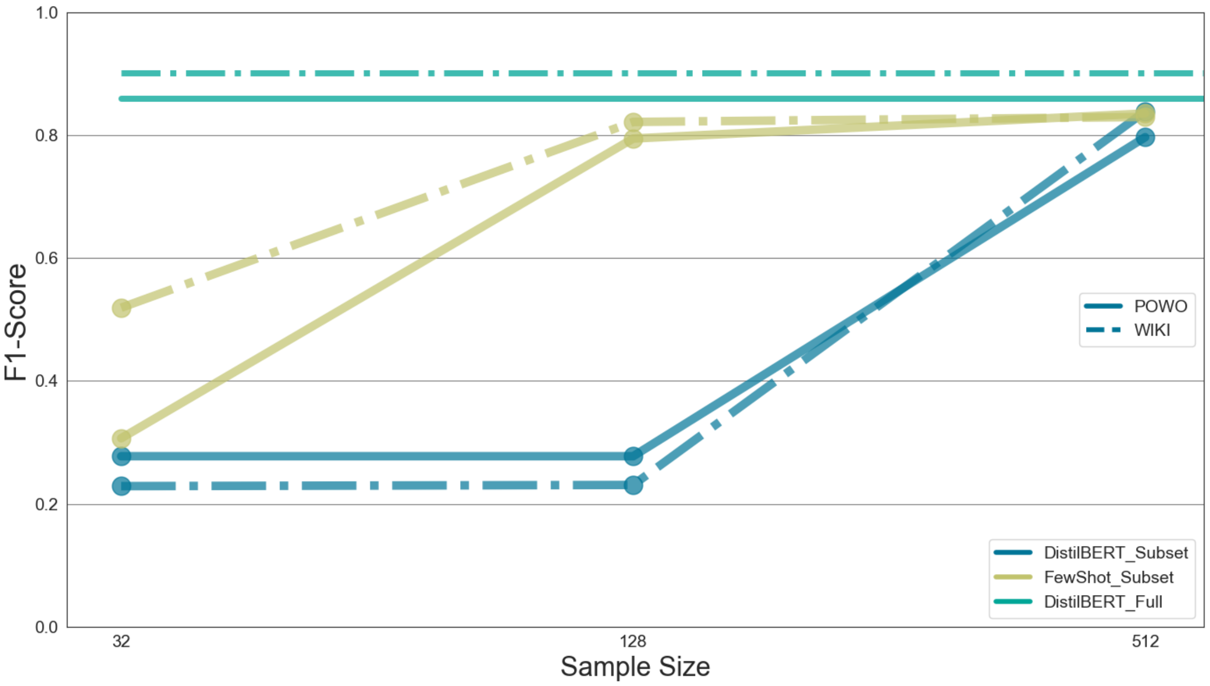

Figure 10: Model results for an evolving sample size for the POWO (solid line) and WIKI (dashed line) datasets. The models are trained on a subset of the original training dataset with a sample size of 32, 128 and 512 and evaluated on the entire dataset of size 1250. We show the results for the standard sequence classification approach using a DistilBERT model (blue) and the few-shot learning approach using an MPNet model (olive). The results when using the entire dataset and DistilBERT model are shown in cyan.

Depending on the specific trait being investigated, we found that 82% to 100% of the original answers contained both a numeric value and a corresponding unit of measurement, making them suitable inputs for the numerical models. On the POWO dataset, the DistilBERT model achieved the lowest average NMAE of 20.42% across all traits. However, there were significant variations in NMAE between traits, with a minimal value of 7.6% for the plant height, 30.5% for the leaf length and 23.2% for the leaf width. In contrast, the NumBERT model, fine-tuned on generated descriptions, exhibited a notably higher NMAE, with values of 44%, 40% and 58.8% on the three traits. Nevertheless, it is important to note that the NumBERT model had, on average, better trait coverage than the DistilBERT model. Specifically, it achieved coverage rates of 47% for plant height, 43% for leaf length and 53% for leaf width, while the DistilBERT model had coverage rates of 27%, 28% and 61% for the same traits. These results indicate that the NumBERT model may predict the same values from the DistilBERT model and potentially more, but it may also more false positives, leading to higher recall at the expense of precision. Therefore, this fine-tuning strategy can prove beneficial as it may increase the performance of the model in an extractive question answering task like this, which has a narrow focus on extracting relevant numeric measurements for each trait. Therefore, with further contrastive training and more negative examples, the NumBERT model could potentially improve its ability to distinguish which numeric measurements are relevant for specific traits and recognize instances where no such measurements exist in a given description.

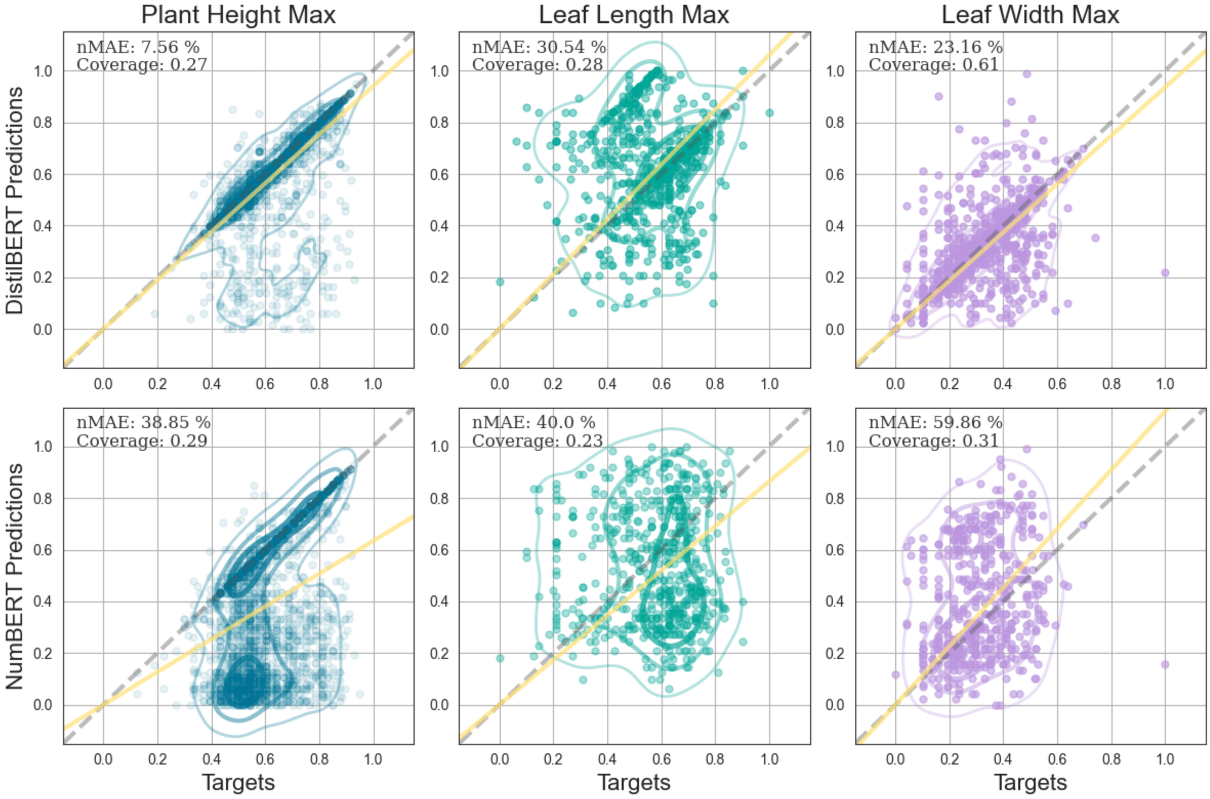

Figure 11: Scatterplot of the true and predicted numerical traits for the DistilBERT (top) and NumBERT (bottom) model on the POWO dataset. The numerical traits are represented as plant height (blue), leaf length (cyan) and leaf width (violet). The 95% and 50% kernel density estimates are also shown by the corresponding trait color. The 1:1 line (gray dashed) and the regression line between the targets and predictions (yellow solid) is also shown

Table 2: The best performing model and its corresponding scores for each NLP task and dataset. For cases when applicable, such as in the topic modelling and categorical trait tasks, where there are several sub-datasets based on input or trait, we show the average per dataset.

Task	**Dataset**	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**NMAE**	**ROUGELSUM**	**Best Model**
Topic Modeling	PREDICTS	0.977	0.885	0.916	0.9	N/A	N/A	EcoBERT DistilBERT
Topic Modeling	LPD	0.968	0.885	0.862	0.872	N/A	N/A	EcoBERT DistilBERT
Text Summarization	bioaRxiv	N/A	N/A	N/A	N/A	N/A	32.83	BART
Text Summarization	PREDICTS	N/A	N/A	N/A	N/A	N/A	36.98	BART
Text Summarization	LPD	N/A	N/A	N/A	N/A	N/A	37.54	BART
Named Entity Recognition	COPIOUS	0.984	0.833	0.787	0.809	N/A	N/A	DeBERTaV3
Named Entity Recognition	SPECIES	0.978	0.79	0.673	0.727	N/A	N/A	DeBERTaV3
Family Classification	POWO	0.968	0.968	0.968	0.968	N/A	N/A	DistilBERT
Family Classification	WIKI	0.971	0.971	0.971	0.97	N/A	N/A	ELECTRA
English Categorical Trait Prediction	POWO	0.912	0.871	0.865	0.866	N/A	N/A	EcoBERT
English Categorical Trait Prediction	WIKI	0.868	0.827	0.811	0.818	N/A	N/A	EcoBERT
Non-English Categorical Trait Prediction	WIKI_ES	0.801	0.752	0.742	0.746	N/A	N/A	Spanish-BERT
Non-English Categorical Trait Prediction	WIKI_DE	0.886	0.869	0.836	0.846	N/A	N/A	German-BERT
Data-Defficient Categorical Trait Prediction	POWO (128 samples)	/	/	/	0.795	N/A	N/A	FewShot
Data-Defficient Categorical Trait Prediction	WIKI (128 samples)	/	/	/	0.822	N/A	N/A	FewShot
Numerical Trait Extraction	POWO	N/A	N/A	N/A	N/A	0.204	N/A	DistilBERT

7 Discussion

Large Language Models present an unprecedented potential for revolutionizing knowledge discovery within the realm of ecological texts. These models showcase remarkable capabilities in understanding and processing the intricacies of ecological language, ranging from taxonomic descriptions to complex ecosystem interactions. Their ability to capture context, comprehend domain-specific jargon, and discern nuanced relationships within ecological texts positions LLMs as invaluable tools for researchers navigating the vast corpus of ecological literature. When it comes to historical texts, LLMs can revolutionize the way researchers delve into ecological archives. By automating literature reviews of centuries-old publications, these models can unearth forgotten ecological knowledge, enabling a deeper understanding of how ecosystems have evolved over time. Ecologists can use LLMs to extract and organize information on species distributions and environmental changes from historical records, shedding light on the historical context of ecological phenomena and facilitating cross-temporal comparisons. Moreover, the application of LLMs extends seamlessly to newly created texts, including research papers and reports. When dealing with contemporary ecological literature, LLMs excel in rapid information retrieval and summarization. They offer the ability to automatically generate concise yet comprehensive summaries of the latest research findings, ensuring ecologists stay abreast of the ever-evolving scientific landscape. Additionally, LLMs can efficiently extract and categorize data within newly created documents, supporting the compilation of comprehensive datasets from contemporary studies. In essence, LLMs serve as versatile tools, bridging the gap between historical ecological texts and the dynamic world of newly generated literature. They empower ecologists to navigate the rich tapestry of ecological knowledge across time, fostering insights, and advancements in the science of ecology.

Within the thesis, we empirically showed the advantages of LLMs over other machine learning and deep learning models across a spectrum of tasks in ecological text analysis. In summary, the performance of LLMs performed across all tasks was remarkable, consistently surpassing the performance of baseline models. Specifically, in tasks related to literature review, the EcoBERT and DistilBERT models achieved remarkable F1-scores exceeding 88% when utilizing paper titles and exceeding 95% when using abstracts. The BART model also delivered impressive results with ROUGE-L-SUM F1-scores exceeding 32% on all three summarization datasets. These results can be used to lower the amount of text that is manually evaluated by researchers or to efficiently create a subset of the text to be used as input in other NLP tasks, lowering the necessary computational power while maintaining the majority of relevant information. In entity extraction tasks, the DeBERTaV3 model stood out, achieving the highest F1-scores of 72.6% and 80.9% on the two datasets, while all LLMs achieved outstanding scores in the family classification task, consistently exceeding 95%, reflecting the discriminative ability of NLP models to pick up on nuances in morphological, anatomical and taxonomical features between plant families. Across the various trait-related tasks, LLMs again demonstrated both excellence and adaptability. Notably, the EcoBERT model exhibited the highest performance when predicting categorical traits from English descriptions, achieving an average F1-score of 83.5%. When applied to Spanish and German descriptions, the monolingual Spanish and German BERT models delivered strong performance, with scores around 75% and 85%, respectively, significantly outperforming the default English-based DistilBERT model. In the data-deficient scenario, the few-shot learning approach proved highly effective, producing competitive results with an average F1-score of 81%, even with a training dataset consisting of only 128 labeled descriptions. Finally, the DistilBERT model excelled in the numerical trait prediction task, achieving the lowest NMAE of 20.6%, averaged across all traits. Overall, these findings underscore the remarkable capabilities of LLMs in various ecological natural language processing tasks. These diverse capabilities can not only expedite ecological research but also offer novel avenues for scientific inquiry and data-driven decision-making. The remarkable performance across these tasks positions LLMs as the state-of-the-art solution for ecological text analysis, showcasing their potential to elevate the efficiency and accuracy of information extraction within the field.

While not of focus in the thesis, it is crucial to recognize that decoders also hold immense potential. With their capacity for text generation and sequence-to-sequence tasks, decoders offer a versatile toolkit for ecological NLP. By harnessing the generative power of decoders, researchers can automatically distill complex ecological insights into easily digestible formats, facilitating faster literature reviews and knowledge synthesis. Furthermore, as NLP models continue to evolve, embracing models with larger parameter sizes becomes increasingly relevant in ecological applications. Models with expanded parameters may exhibit a more profound understanding of ecological nuances, as their enhanced capacity enables them to capture intricate patterns and relationships within texts [130]. This depth of comprehension proves invaluable when extracting ecological information, categorizing species, predicting traits, or summarizing complex findings. As computational resources become more accessible, leveraging larger models allows ecologists to push the boundaries of what is achievable in ecological NLP, unlocking new possibilities for understanding and conserving our natural world. While we focused on only the base version of the models, most of them also contain different checkpoints with a larger parameter set and trained on bigger datasets.

However, the adoption of LLMs in ecological applications is not without its challenges and limitations. One prominent concern revolves around the interpretability of these models [41, 70]. While LLMs exhibit remarkable performance, understanding the underlying decision-making processes can be challenging, raising questions about the reliability and trustworthiness of their outputs [124], especially in sensitive ecological decision-making scenarios. This lack of interpretability can hinder effective collaboration between ecologists and computer scientists, as it becomes challenging to explain the model's rationale and decision criteria. Additionally, the substantial computational resources required to train and fine-tune LLMs pose barriers for researchers with limited access to high-performance computing infrastructure. Bridging this resource gap and fostering collaboration between ecological researchers and computational experts is imperative to ensure that the benefits of LLMs are accessible to the broader ecological community. Furthermore, LLMs may exhibit biases present in the training data, potentially perpetuating and amplifying existing biases in ecological knowledge [36, 116]. These biases can be particularly problematic when LLMs are used to inform ecological decision-making or policy recommendations, as they may inadvertently reinforce unfair or skewed perspectives. Addressing these biases requires meticulous data curation, transparent model evaluation, and continuous refinement of model training processes. Additionally, interdisciplinary collaboration between ecologists and computer scientists is essential to develop bias mitigation strategies tailored to the unique challenges of ecological data. As we leverage the power of LLMs in ecological research, it is crucial to address these limitations systematically, ensuring responsible and ethical use to harness the full potential of these models for advancing ecological science while fostering productive collaborations that bridge disciplinary boundaries.

To advance the field of ecological NLP and fully leverage the potential of LLMs, the development of new benchmark datasets tailored to ecological tasks is imperative. These datasets should encompass a range of tasks, including sequence classification, token classification, text summarization, and question answering, to comprehensively evaluate the capabilities of NLP models within the ecological context. While creating entirely new datasets is one avenue, it is equally essential to explore innovative approaches for dataset creation. One promising approach involves harnessing existing ecological texts, a treasure trove of information, and combining them with labels from diverse sources. This strategy allows us to repurpose and augment existing ecological literature, efficiently transforming it into labeled datasets suitable for training and evaluating NLP models. By employing such creative solutions, we can tackle the scarcity of labeled ecological data, paving the way for more robust, specialized, and ecologically relevant benchmark datasets. These efforts will not only bolster the performance of NLP models but also foster the growth of ecological NLP as a discipline, empowering researchers to tackle pressing ecological challenges more effectively and comprehensively.

8 Conclusion & outlook

In conclusion, large language models (LLMs) represent a game-changing force in ecological knowledge discovery. Their proficiency in comprehending the intricacies of ecological texts, coupled with their unmatched performance in text summarization, classification, and question answering, positions them as indispensable tools for ecologists and researchers. By streamlining the process of literature review, automating information extraction, and enhancing data-driven decision-making, LLMs have the potential to expedite ecological research and foster deeper insights into the complex web of relationships within ecosystems. However, it is essential to navigate the ethical and interpretability challenges that come with harnessing the power of these models responsibly.

Looking ahead, the future of ecological research is poised for a synergy between LLMs and other cutting-edge technologies. The integration of LLMs with computer vision models can pave the way for holistic ecological studies by combining textual data with image and sensor data, offering a more comprehensive understanding of ecosystems. Furthermore, collaborative efforts between LLMs and domain-specific ecological models can lead to hybrid models that leverage the strengths of both to address specific ecological challenges.

This area of research holds significant promise, especially with LLMs pre-trained from scratch on extensive ecological text corpora. Such models have the potential to tailor their vocabularies to the ecological domain, leading to substantial improvements in their reasoning abilities for ecological tasks. While our EcoBERT model provides a glimpse of this potential, it is worth noting that it retained its original pre-training on a general English corpus and only received additional pre-training on a relatively small ecological dataset. To fully harness this potential, expanding the training corpus to encompass larger databases, such as the texts from the Biodiversity Heritage Library, holds great promise. However, in the rapidly evolving landscape of AI and NLP research, there is a potential challenge. Even after substantial investments in training such models, they may be surpassed by general-domain language models that draw from much larger resources and novel methodologies. This dynamic raises important questions about the future direction of ecological NLP models. To navigate this evolving landscape effectively, we can draw valuable insights from other fields that face similar challenges and adaptations in the face of rapid technological advancements. These insights can help guide the development of ecological NLP models and ensure their continued relevance and impact in the field of ecology. This interdisciplinary approach holds the promise of not only improving ecological data analysis but also enabling novel discoveries and innovative solutions for pressing environmental issues. As we move forward, embracing the capabilities of LLMs and their integration with diverse modeling approaches is likely to shape the future of ecological research, driving the field toward greater depth and efficiency in uncovering the mysteries of our natural world.

Declaration of authorship

I hereby assure that this thesis was exclusively made by myself and that I have used no other sources and aids than the ones enlisted.

Viktor Domazetoski
31.01.2024

References

[1] Majid Afshar, Dmitriy Dligach, Brihat Sharma, Xiaoyuan Cai, Jason Boyda, Steven Birch, Daniel Valdez, Suzan Zelisko, Cara Joyce, François Modave, et al. Development and application of a high throughput natural language processing architecture to convert all clinical documents in a clinical data warehouse into standardized medical vocabularies. *Journal of the American Medical Informatics Association*, 26(11):1364–1369, 2019.

[2] Mohamed AlShuweihi, Said A Salloum, and Khaled Shaalan. Biomedical corpora and natural language processing on clinical text in languages other than english: a systematic review. *Recent advances in intelligent systems and smart applications*, pages 491–509, 2021.

[3] Phillips R. Milo R. Bar-On, Y. M. The biomass distribution on earth. *Proceedings of the National Academy of Sciences, 115(25), 6506-6511.*, 2018.

[4] Jan Beck, Marianne Böller, Andreas Erhardt, and Wolfgang Schwanghart. Spatial bias in the gbif database and its effect on modeling species' geographic distributions. *Ecological Informatics*, 19:10–15, 2014.

[5] Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*, 2019.

[6] Daniel Bikel and Imed Zitouni. *Multilingual natural language processing applications: from theory to practice.* IBM Press, 2012.

[7] Lirong Cai, Holger Kreft, Amanda Taylor, Pierre Denelle, Julian Schrader, Franz Essl, Mark van Kleunen, Jan Pergl, Petr Pyšek, Anke Stein, et al. Global models and predictions of plant diversity based on advanced machine learning techniques. *New Phytologist*, 237(4):1432–1445, 2023.

[8] Bradley J Cardinale, J Emmett Duffy, Andrew Gonzalez, David U Hooper, Charles Perrings, Patrick Venail, Anita Narwani, Georgina M Mace, David Tilman, David A Wardle, et al. Biodiversity loss and its impact on humanity. *Nature*, 486(7401):59–67, 2012.

[9] Jeannine Cavender-Bares, Walter Jetz, Ryan Pavlick, David Schimel, John Arthur Gamon, Sarah E Hobbie, and Philip A Townsend. A global remote sensing mission to detect and predict plant functional biodiversity change. In *AGU Fall Meeting Abstracts*, volume 2015, pages B43K–05, 2015.

[10] José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*, 2020.

[11] Scott Chamberlain, Eduard Szoecs, Zachary Foster, Zebulun Arendsee, Carl Boettiger, Karthik Ram, Ignasi Bartomeus, John Baumgartner, James O'Donnell, Jari

Oksanen, et al. taxize: Taxonomic information from around the web. *R package version 0.9*, 92, 2020.

[12] Scott A Chamberlain and Eduard Szöcs. taxize: taxonomic search and retrieval in r. *F1000Research*, 2, 2013.

[13] Branden Chan, Stefan Schweter, and Timo Möller. German's next language model. *arXiv preprint arXiv:2010.10906*, 2020.

[14] Andrea Chaves, Cyrille Kesiku, and Begonya Garcia-Zapirain. Automatic text summarization of biomedical text data: A systematic review. *Information*, 13(8):393, 2022.

[15] Eya Cherif, Hannes Feilhauer, Katja Berger, Phuong D Dao, Michael Ewald, Tobias B Hank, Yuhong He, Kyle R Kovach, Bing Lu, Philip A Townsend, et al. From spectra to plant functional traits: Transferable multi-trait models from heterogeneous and sparse data. *Remote Sensing of Environment*, 292:113580, 2023.

[16] KR1442 Chowdhary and KR Chowdhary. Natural language processing. *Fundamentals of artificial intelligence*, pages 603–649, 2020.

[17] Maarten JM Christenhusz and James W Byng. The number of known plants species in the world and its annual increase. *Phytotaxa*, 261(3):201–217, 2016.

[18] Sylvain Christin, Éric Hervet, and Nicolas Lecomte. Applications for deep learning in ecology. *Methods in Ecology and Evolution*, 10(10):1632–1644, 2019.

[19] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

[20] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models, 2022.

[21] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*, 2020.

[22] Elijah Cole, Grant Van Horn, Christian Lange, Alexander Shepard, Patrick Leary, Pietro Perona, Scott Loarie, and Oisin Mac Aodha. Spatial implicit neural representations for global-scale species mapping. *arXiv preprint arXiv:2306.02564*, 2023.

[23] Ben Collen, Mala Ram, Tara Zamin, and Louise McRae. The tropical biodiversity data gap: addressing disparity in global monitoring. *Tropical Conservation Science*, 1(2):75–88, 2008.

[24] Richard Cornford, Stefanie Deinet, Adriana De Palma, Samantha LL Hill, Louise McRae, Benjamin Pettit, Valentina Marconi, Andy Purvis, and Robin Freeman. Fast, scalable, and automated identification of articles for biodiversity and macroecological datasets. *Global Ecology and Biogeography*, 30(1):339–347, 2021.

[25] Robert H Cowie, Philippe Bouchet, and Benoît Fontaine. The sixth mass extinction: fact, fiction or speculation? *Biological Reviews*, 97(2):640–663, 2022.

[26] Biraj Dahal, Sathish AP Kumar, and Zhenlong Li. Topic modeling and sentiment analysis of global climate change tweets. *Social network analysis and mining*, 9:1–20, 2019.

[27] Pierre Denelle, Patrick Weigelt, and Holger Kreft. Gift an r package to access the global inventory of floras and traits. *bioRxiv*, pages 2023–06, 2023.

[28] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[29] Sandra Díaz, Jens Kattge, Johannes HC Cornelissen, Ian J Wright, Sandra Lavorel, Stéphane Dray, Björn Reu, Michael Kleyer, Christian Wirth, I Colin Prentice, et al. The global spectrum of plant form and function. *Nature*, 529(7585):167–171, 2016.

[30] Viktor Domazetoski, Holger Kreft, Helena Bestova, Philipp Wieder, Radoslav Koynov, Alireza Zarei, and Patrick Weigelt. Using natural language processing to extract plant functional traits from unstructured text. *Unpublished Manuscript*, 2023.

[31] Yigal Elad and Ilaria Pertot. Climate change impacts on plant pathogens and plant diseases. *Journal of Crop Improvement*, 28(1):99–139, 2014.

[32] Lorena Endara, Hong Cui, and J Gordon Burleigh. Extraction of phenotypic traits from taxonomic descriptions for the tree of life using natural language processing. *Applications in Plant Sciences*, 6(3):e1035, 2018.

[33] Daniel Falster, Rachael Gallagher, Elizabeth H Wenk, Ian J Wright, Dony Indiarto, Samuel C Andrew, Caitlan Baxter, James Lawson, Stuart Allen, Anne Fuchs, et al. Austraits, a curated plant trait database for the australian flora. *Scientific Data*, 8(1):254, 2021.

[34] Scott S Farley, Andria Dawson, Simon J Goring, and John W Williams. Situating ecology as a big-data science: Current advances, challenges, and solutions. *BioScience*, 68(8):563–576, 2018.

[35] Maxwell J Farrell, Liam Brierley, Anna Willoughby, Andrew Yates, and Nicole Mideo. Past and future uses of text mining in ecology and evolution. *Proceedings of the Royal Society B*, 289(1975):20212721, 2022.

[36] Emilio Ferrara. Should chatgpt be biased? challenges and risks of bias in large language models. *arXiv preprint arXiv:2304.03738*, 2023.

[37] Dilek Fraisl, Gerid Hager, Baptiste Bedessem, Margaret Gold, Pen-Yuan Hsing, Finn Danielsen, Colleen B Hitchcock, Joseph M Hulbert, Jaume Piera, Helen Spiers, et al. Citizen science in environmental and ecological sciences. *Nature Reviews Methods Primers*, 2(1):64, 2022.

[38] Janet Franklin, Josep M Serra-Diaz, Alexandra D Syphard, and Helen M Regan. Big data for forecasting the impacts of global change on plant communities. *Global Ecology and Biogeography*, 26(1):6–17, 2017.

[39] Rachael V Gallagher, Daniel S Falster, Brian S Maitner, Roberto Salguero-Gómez, Vigdis Vandvik, William D Pearse, Florian D Schneider, Jens Kattge, Jorrit H Poelen, Joshua S Madin, et al. Open science principles for accelerating trait-based science across the tree of life. *Nature ecology & evolution*, 4(3):294–303, 2020.

[40] Martin Gerner, Goran Nenadic, and Casey M Bergman. Linnaeus: a species name identification system for biomedical literature. *BMC bioinformatics*, 11(1):1–17, 2010.

[41] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE, 2018.

[42] John M Giorgi and Gary D Bader. Transfer learning for biomedical named entity recognition with neural networks. *Bioinformatics*, 34(23):4087–4094, 2018.

[43] ISSG GISD. Global invasive species database. 2020.

[44] Hervé Goëau, Pierre Bonnet, Alexis Joly, Vera Bakić, Julien Barbe, Itheri Yahiaoui, Souheil Selmi, Jennifer Carré, Daniel Barthélémy, Nozha Boujemaa, et al. Pl@ ntnet mobile app. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 423–424, 2013.

[45] Rafaël Govaerts, Eimear Nic Lughadha, Nicholas Black, Robert Turner, and Alan Paton. The world checklist of vascular plants, a continuously updated resource for exploring global plant diversity. *Scientific Data*, 8(1):215, 2021.

[46] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23, 2021.

[47] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*, 2020.

[48] Nancy E Gwinn and Constance Rinaldo. The biodiversity heritage library: sharing biodiversity literature with the world. *IFLA journal*, 35(1):25–34, 2009.

[49] Charles R Harris, K Jarrod Millman, Stéfan J Van Der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J Smith, et al. Array programming with numpy. *Nature*, 585(7825):357–362, 2020.

[50] Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*, 2021.

[51] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*, 2020.

[52] Jonathan J Henn, Vanessa Buzzard, Brian J Enquist, Aud H Halbritter, Kari Klanderud, Brian S Maitner, Sean T Michaletz, Christine Pötsch, Lorah Seltzer, Richard J Telford, et al. Intraspecific trait variation and phenotypic plasticity mediate alpine plant species response to climate change. *Frontiers in Plant Science*, 9:1548, 2018.

[53] Julia Hirschberg and Christopher D Manning. Advances in natural language processing. *Science*, 349(6245):261–266, 2015.

[54] Lawrence N Hudson, Tim Newbold, Sara Contu, Samantha LL Hill, Igor Lysenko, Adriana De Palma, Helen RP Phillips, Rebecca A Senior, Dominic J Bennett, Hollie Booth, et al. The predicts database: a global database of how local terrestrial biodiversity responds to human impacts. *Ecology and evolution*, 4(24):4701–4735, 2014.

[55] John D Hunter. Matplotlib: A 2d graphics environment. *Computing in science & engineering*, 9(03):90–95, 2007.

[56] NV Ivanova and MP Shashkov. The possibilities of gbif data use in ecological research. *Russian Journal of Ecology*, 52:1–8, 2021.

[57] Jens Kattge, Gerhard Bönisch, Sandra Díaz, Sandra Lavorel, Iain Colin Prentice, Paul Leadley, Susanne Tautenhahn, Gijsbert DA Werner, Tuomas Aakala, Mehdi Abedi, et al. Try plant trait database–enhanced coverage and open access. *Global change biology*, 26(1):119–188, 2020.

[58] Jens Kattge, Sandra Diaz, Sandra Lavorel, I Colin Prentice, Paul Leadley, Gerhard Bönisch, Eric Garnier, Mark Westoby, Peter B Reich, Ian J Wright, et al. Try–a global database of plant traits. *Global change biology*, 17(9):2905–2935, 2011.

[59] Gunnar Keppel, Dylan Craven, Patrick Weigelt, Stephen A Smith, Masha T van der Sande, Brody Sandel, Sam C Levin, Holger Kreft, and Tiffany M Knight. Synthesizing tree biodiversity data to understand global patterns and processes of vegetation. *Journal of Vegetation Science*, 32(3):e13021, 2021.

[60] Christian König, Patrick Weigelt, Julian Schrader, Amanda Taylor, Jens Kattge, and Holger Kreft. Biodiversity data integration—the significance of data resolution and domain. *PLoS biology*, 17(3):e3000183, 2019.

[61] Drew Koning, Indra Neil Sarkar, and Thomas Moritz. Taxongrab: Extracting taxonomic names from text. *Biodiversity Informatics*, 2:79–82, 2005.

[62] KO Konno, Munemitsu Akasaka, Chieko Koshida, Naoki Katayama, Noriyuki Osada, Rebecca Spake, and Tatsuya Amano. Ignoring non-english-language studies may bias ecological meta-analyses. *Ecology and Evolution*, 10(13):6373–6384, 2020.

[63] Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*, 2018.

[64] Armelle Lajeunesse and Yoan Fourcade. Temporal analysis of gbif data reveals the restructuring of communities following climate change. *Journal of Animal Ecology*, 92(2):391–402, 2023.

[65] Nicolas Le Guillarme and Wilfried Thuiller. Taxonerd: deep neural models for the recognition of taxonomic entities in the ecological and evolutionary literature. *Methods in Ecology and Evolution*, 13(3):625–641, 2022.

[66] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[67] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

[68] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461, 2019.

[69] C Lin. Recall-oriented understudy for gisting evaluation (rouge). *Retrieved August*, 20:2005, 2005.

[70] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1):18, 2020.

[71] Lin Liu, Lin Tang, Wen Dong, Shaowen Yao, and Wei Zhou. An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus*, 5(1):1–22, 2016.

[72] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[73] Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li, and Jun Zhao. Finbert: A pretrained financial language representation model for financial text mining. In *Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence*, pages 4513–4519, 2021.

[74] Edward Loper and Steven Bird. Nltk: The natural language toolkit. *arXiv preprint cs/0205028*, 2002.

[75] Marc Moreno Lopez and Jugal Kalita. Deep learning applied to nlp. *arXiv preprint arXiv:1703.03091*, 2017.

[76] Eimear Nic Lughadha, Rafael Govaerts, Irina Belyaeva, Nicholas Black, Heather Lindon, ROBERT ALLKIN, ROBERT E MAGILL, and Nicky Nicolson. Counting counts: revised estimates of numbers of accepted species of flowering plants, seed plants, vascular plants and land plants with a review of other recent estimates. *Phytotaxa*, 272(1):82–88, 2016.

[77] Jianxia Ma and Hui Yuan. Bi-lstm+ crf-based named entity recognition in scientific papers in the field of ecological restoration technology. *Proceedings of the Association for Information Science and Technology*, 56(1):186–195, 2019.

[78] Patrick Mäder, David Boho, Michael Rzanny, Marco Seeland, Hans Christian Wittich, Alice Deggelmann, and Jana Wäldchen. The flora incognita app–interactive plant species identification. *Methods in Ecology and Evolution*, 12(7):1335–1342, 2021.

[79] Brian S Maitner, Brad Boyle, Nathan Casler, Rick Condit, John Donoghue, Sandra M Durán, Daniel Guaderrama, Cody E Hinchliff, Peter M Jørgensen, Nathan JB Kraft, et al. The bien r package: A tool to access the botanical information and ecology network (bien) database. *Methods in Ecology and Evolution*, 9(2):373–379, 2018.

[80] Brian S Maitner, Rachael Gallagher, Jens-Christian Svenning, Melanie Tietje, Elizabeth H Wenk, and Wolf L Eiserhardt. Socioeconomics and biogeography jointly drive geographic biases in our knowledge of plant traits: a global assessment of the raunkiaerian shortfall in plants. *bioRxiv*, pages 2022–09, 2022.

[81] Emily McCallen, Jonathan Knott, Gabriela Nunez-Mir, Benjamin Taylor, Insu Jo, and Songlin Fei. Trends in ecology: shifts in ecological research themes over the past four decades. *Frontiers in Ecology and the Environment*, 17(2):109–116, 2019.

[82] Wes McKinney et al. pandas: a foundational python library for data analysis and statistics. *Python for high performance and scientific computing*, 14(9):1–9, 2011.

[83] Carsten Meyer, Patrick Weigelt, and Holger Kreft. Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecology letters*, 19(8):992–1006, 2016.

[84] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[85] Angela T Moles, David D Ackerly, John C Tweddle, John B Dickie, Roger Smith, Michelle R Leishman, Margaret M Mayfield, Andy Pitman, Jeff T Wood, and Mark Westoby. Global patterns in seed size. *Global ecology and biogeography*, 16(1):109–116, 2007.

[86] Maria Auxiliadora Mora and José Enrique Araya. Semi-automatic extraction of plants morphological characters from taxonomic descriptions written in spanish. *Biodiversity data journal*, (6), 2018.

[87] Gil Nelson and Shari Ellis. The history and impact of digitization and digital data mobilization on biodiversity research. *Philosophical Transactions of the Royal Society B*, 374(1763):20170391, 2019.

[88] Nhung TH Nguyen, Roselyn S Gabud, and Sophia Ananiadou. Copious: A gold standard corpus of named entities towards extracting species occurrence from biodiversity literature. *Biodiversity data journal*, (7), 2019.

[89] Evangelos Pafilis, Sune P Frankild, Lucia Fanini, Sarah Faulwetter, Christina Pavloudi, Aikaterini Vasileiadou, Christos Arvanitidis, and Lars Juhl Jensen. The species and organisms resources for fast and accurate identification of taxonomic names in text. *PloS one*, 8(6):e65390, 2013.

[90] Camille Parmesan and Mick E Hanley. Plants and climate change: complexities and surprises. *Annals of botany*, 116(6):849–864, 2015.

[91] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[92] Maximilian Pichler and Florian Hartig. Machine learning and deep learning—a review for ecologists. *Methods in Ecology and Evolution*, 14(4):994–1016, 2023.

[93] Stephen M Powers and Stephanie E Hampton. Open science, reproducibility, and transparency in ecology. *Ecological applications*, 29(1):e01822, 2019.

[94] Andy Purvis, Tim Newbold, Adriana De Palma, Sara Contu, Samantha LL Hill, Katia Sanchez-Ortiz, Helen RP Phillips, Lawrence N Hudson, Igor Lysenko, Luca Börger, et al. Modelling and projecting the response of local terrestrial biodiversity worldwide to land use and related pressures: the predicts project. In *Advances in ecological research*, volume 58, pages 201–241. Elsevier, 2018.

[95] Petr Pyšek, David M Richardson, Jan Pergl, Vojtěch Jarošík, Zuzana Sixtová, and Ewald Weber. Geographical and taxonomic biases in invasion ecology. *Trends in ecology & evolution*, 23(5):237–244, 2008.

[96] Hong Qian, Jian Zhang, and Jingchao Zhao. How many known vascular plant species are there in the world? an integration of multiple global plant databases. *Biodiversity Science*, 30(7):22254, 2022.

[97] Sandra Quijas, Bernhard Schmid, and Patricia Balvanera. Plant diversity enhances provision of ecosystem services: A new synthesis. *Basic and Applied Ecology*, 11(7):582–593, 2010.

[98] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[99] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

[100] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.

[101] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[102] Guido Sautter, Klemens Böhm, and Donat Agosti. A combining approach to find all taxon names (fat). *Biodiversity informatics*, 3, 2006.

[103] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

[104] Christopher Schiller, Sebastian Schmidtlein, Coline Boonman, Alvaro Moreno-Martínez, and Teja Kattenborn. Deep learning and citizen science enable automated plant trait predictions from photographs. *Scientific Reports*, 11(1):16395, 2021.

[105] Franziska Schrodt, Jens Kattge, Hanhuai Shan, Farideh Fazayeli, Julia Joswig, Arindam Banerjee, Markus Reichstein, Gerhard Bönisch, Sandra Díaz, John Dickie, et al. Bhpmf–a hierarchical b ayesian approach to gap-filling and trait prediction for macroecology and functional biogeography. *Global Ecology and Biogeography*, 24(12):1510–1521, 2015.

[106] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015.

[107] Alex Sherstinsky. Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *Physica D: Nonlinear Phenomena*, 404:132306, 2020.

[108] Pamela S Soltis. Digitization of herbaria enables novel research. *American journal of botany*, 104(9):1281–1284, 2017.

[109] Bosheng Song, Fen Li, Yuansheng Liu, and Xiangxiang Zeng. Deep learning methods for biomedical named entity recognition: a survey and qualitative comparison. *Briefings in Bioinformatics*, 22(6):bbab282, 2021.

[110] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33:16857–16867, 2020.

[111] Robin Steenweg, Mark Hebblewhite, Roland Kays, Jorge Ahumada, Jason T Fisher, Cole Burton, Susan E Townsend, Chris Carbone, J Marcus Rowcliffe, Jesse Whittington, et al. Scaling-up camera traps: Monitoring the planet's biodiversity with networks of remote sensors. *Frontiers in Ecology and the Environment*, 15(1):26–34, 2017.

[112] Larissa Sayuri Moreira Sugai, Thiago Sanna Freire Silva, José Wagner Ribeiro Jr, and Diego Llusia. Terrestrial passive acoustic monitoring: review and perspectives. *BioScience*, 69(1):15–25, 2019.

[113] Nathan G Swenson. Phylogenetic imputation of plant functional trait databases. *Ecography*, 37(2):105–110, 2014.

[114] Amanda Taylor, Gerhard Zotz, Patrick Weigelt, Lirong Cai, Dirk Nikolaus Karger, Christian König, and Holger Kreft. Vascular epiphytes contribute disproportionately to global centres of plant diversity. *Global Ecology and Biogeography*, 31(1):62–74, 2022.

[115] Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*, 2022.

[116] Surendrabikram Thapa and Surabhi Adhikari. Chatgpt, bard, and large language models for biomedical research: Opportunities and pitfalls. *Annals of Biomedical Engineering*, pages 1–5, 2023.

[117] Amirsina Torfi, Rouzbeh A Shirvani, Yaser Keneshloo, Nader Tavaf, and Edward A Fox. Natural language processing advancements by deep learning: A survey. *arXiv preprint arXiv:2003.01200*, 2020.

[118] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[119] Devis Tuia, Benjamin Kellenberger, Sara Beery, Blair R Costelloe, Silvia Zuffi, Benjamin Risse, Alexander Mathis, Mackenzie W Mathis, Frank van Langevelde, Tilo Burghardt, et al. Perspectives in machine learning for wildlife conservation. *Nature communications*, 13(1):792, 2022.

[120] Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. Efficient few-shot learning without prompts. *arXiv preprint arXiv:2209.11055*, 2022.

[121] Kristin L Vanderbilt, Chau-Chin Lin, Sheng-Shan Lu, Abd Rahman Kassim, Honglin He, Xuebing Guo, Inigo San Gil, David Blankman, and John H Porter. Fostering ecological data sharing: collaborations in the international long term ecological research network. *Ecosphere*, 6(10):1–18, 2015.

[122] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[123] Kris Verheyen, Olivier Honnay, Glenn Motzkin, Martin Hermy, and David R Foster. Response of forest plant species to land-use change: a life-history trait-based approach. *Journal of Ecology*, pages 563–577, 2003.

[124] Kerstin N Vokinger, Stefan Feuerriegel, and Aaron S Kesselheim. Mitigating bias in machine learning for medicine. *Communications medicine*, 1(1):25, 2021.

[125] Jana Wäldchen, Michael Rzanny, Marco Seeland, and Patrick Mäder. Automated plant species identification—trends and future directions. *PLoS computational biology*, 14(4):e1005993, 2018.

[126] Barnaby E Walker, Tarciso CC Leão, Steven P Bachman, Eve Lucas, and Eimear Nic Lughadha. Evidence-based guidelines for automated conservation assessments of plant species. *Conservation Biology*, 37(1):e13992, 2023.

[127] Changhan Wang, Kyunghyun Cho, and Jiatao Gu. Neural machine translation with byte-level subwords. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9154–9160, 2020.

[128] Michael L Waskom. Seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021, 2021.

[129] Brian C Weeks, Zhizhuo Zhou, Bruce K O'Brien, Rachel Darling, Morgan Dean, Tiffany Dias, Gemmechu Hassena, Mingyu Zhang, and David F Fouhey. A deep neural network for high-throughput measurement of functional traits on museum skeletal specimens. *Methods in Ecology and Evolution*, 14(2):347–359, 2023.

[130] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

[131] Patrick Weigelt, Christian König, and Holger Kreft. Gift–a global inventory of floras and traits for macroecology and biogeography. *Journal of Biogeography*, 47(1):16–43, 2020.

[132] David Westergaard, Hans-Henrik Stærfeldt, Christian Tønsberg, Lars Juhl Jensen, and Søren Brunak. A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts. *PLoS computational biology*, 14(2):e1005962, 2018.

[133] Kathy Willis. *State of the world's plants 2017*. Royal Botanics Gardens Kew, 2017.

[134] Sophie Wolf, Miguel D Mahecha, Francesco Maria Sabatini, Christian Wirth, Helge Bruelheide, Jens Kattge, Álvaro Moreno Martínez, Karin Mora, and Teja Kattenborn. Citizen science plant observations encode global trait patterns. *Nature Ecology & Evolution*, 6(12):1850–1859, 2022.

[135] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

[136] T Elizabeth Workman, Marcelo Fiszman, and John F Hurdle. Text summarization as a decision support aid. *BMC medical informatics and decision making*, 12:1–12, 2012.

[137] Ian J Wright, Peter B Reich, Mark Westoby, David D Ackerly, Zdravko Baruch, Frans Bongers, Jeannine Cavender-Bares, Terry Chapin, Johannes HC Cornelissen, Matthias Diemer, et al. The worldwide leaf economics spectrum. *Nature*, 428(6985):821–827, 2004.

[138] Wenjing Yang, Keping Ma, and Holger Kreft. Geographical sampling bias in a large distributional database and its effects on species richness–environment models. *Journal of Biogeography*, 40(8):1415–1426, 2013.

Appendix

Table 3: Model Results on the Topic Modeling Task.

Dataset	Text	Accuracy	Precision	Recall	F1-Score	Model
LPI	Title	93.258	70.673	81.216	75.578	Logistic Regression
LPI	Title	95.813	87.195	79.006	82.899	EcoBERT
LPI	Title	95.813	87.195	79.006	82.899	DistilBERT
LPI	Title	95.884	87.730	79.006	83.140	DeBERTaV3
LPI	Title	95.884	85.965	81.215	83.523	ELECTRA
LPI	Abstract	96.026	80.193	91.713	85.567	Logistic Regression
LPI	Abstract	97.800	89.894	93.370	91.599	EcoBERT
LPI	Abstract	97.800	89.894	93.370	91.599	DistilBERT
LPI	Abstract	97.729	91.620	90.608	91.111	DeBERTaV3
LPI	Abstract	97.516	89.674	91.160	90.411	ELECTRA
PREDICTS	Title	96.171	78.022	91.613	84.273	Logistic Regression
PREDICTS	Title	96.965	85.093	88.387	86.709	EcoBERT
PREDICTS	Title	96.965	85.093	88.387	86.709	DistilBERT
PREDICTS	Title	96.532	82.036	88.387	85.093	DeBERTaV3
PREDICTS	Title	96.532	84.967	83.871	84.416	ELECTRA
PREDICTS	Abstract	97.327	84.302	93.548	88.685	Logistic Regression
PREDICTS	Abstract	98.483	91.875	94.839	93.333	EcoBERT
PREDICTS	Abstract	98.483	91.875	94.839	93.333	DistilBERT
PREDICTS	Abstract	97.977	88.957	93.548	91.195	DeBERTaV3
PREDICTS	Abstract	98.049	89.506	93.548	91.483	ELECTRA

Table 4: Model Results on the Text Summarization Task.

Dataset	Rouge1	Rouge2	RougeL	RougeLsum	Model
bioarXiv	18.550	6.380	13.620	15.660	Baseline
bioarXiv	37.643	15.582	31.302	31.360	FLAN-T5
bioarXiv	38.836	16.743	32.800	32.830	BART
LPD	11.210	5.230	8.680	8.670	Baseline
LPD	40.655	19.503	35.157	35.116	FLAN-T5
LPD	42.828	22.166	37.599	37.541	BART
PREDICTS	11.180	5.090	8.650	8.650	Baseline
PREDICTS	40.444	18.348	34.478	34.497	FLAN-T5
PREDICTS	42.792	20.937	36.979	36.982	BART

Table 5: Model Results on the Named Entity Recognition Task.

Dataset	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**Model**
COPIOUS	97.475	66.957	61.436	64.078	EcoBERT
COPIOUS	97.523	65.652	60.561	63.004	DistilBERT
COPIOUS	98.418	83.333	78.674	80.936	DeBERTaV3
COPIOUS	98.143	76.812	73.611	75.177	ELECTRA
S800	96.982	61.290	52.778	56.716	EcoBERT
S800	96.890	61.290	53.293	57.012	DistilBERT
S800	97.831	78.992	67.303	72.680	DeBERTaV3
S800	97.579	71.809	61.465	66.235	ELECTRA

Table 6: Model Results on the Family Classification Task.

Dataset	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**Model**
POWO	96.240	96.322	96.334	96.314	Logistic Regression
POWO	96.480	96.542	96.538	96.521	EcoBERT
POWO	96.800	96.840	96.815	96.823	DistilBERT
POWO	94.800	94.973	94.870	94.894	DeBERTa_v3
POWO	96.400	96.510	96.404	96.437	ELECTRA
WIKI	96.160	96.133	96.163	96.134	Logistic Regression
WIKI	97.120	97.069	97.015	97.025	EcoBERT
WIKI	97.120	97.070	97.007	97.028	DistilBERT
WIKI	96.560	96.552	96.446	96.473	DeBERTa_v3
WIKI	97.120	97.061	97.056	97.030	ELECTRA

Table 7: Model Results on the Categorical Trait Classification of English Descriptions Task.

Dataset	Trait	Accuracy	Precision	Recall	F1-Score	Model
POWO	Growth Form	77.573	76.708	52.799	60.532	Keyword Search
POWO	Growth Form	88.400	79.509	79.081	79.276	Logistic Regression
POWO	Growth Form	92.400	87.793	85.484	86.379	EcoBERT
POWO	Growth Form	92.240	86.182	85.943	86.062	DistilBERT
POWO	Growth Form	91.520	85.038	86.292	85.575	DeBERTav3
POWO	Growth Form	91.760	87.252	83.332	84.952	ELECTRA
POWO	Life Form	80.000	0.000	0.000	0.000	Keyword Search
POWO	Life Form	86.640	82.813	81.278	81.939	Logistic Regression
POWO	Life Form	90.000	86.387	87.472	86.902	EcoBERT
POWO	Life Form	89.760	85.866	87.624	86.613	DistilBERT
POWO	Life Form	85.760	78.926	80.188	79.480	DeBERTav3
POWO	Life Form	87.520	81.859	83.028	82.410	ELECTRA
WIKI	Growth Form	78.293	86.644	48.078	59.151	Keyword Search
WIKI	Growth Form	81.920	78.838	79.678	79.218	Logistic Regression
WIKI	Growth Form	90.800	89.753	88.950	89.331	EcoBERT
WIKI	Growth Form	91.360	90.292	89.931	90.094	DistilBERT
WIKI	Growth Form	90.480	89.745	87.995	88.752	DeBERTav3
WIKI	Growth Form	89.600	88.194	87.822	88.001	ELECTRA
WIKI	Life Form	80.032	40.000	0.325	0.644	Keyword Search
WIKI	Life Form	73.520	62.395	61.558	61.921	Logistic Regression
WIKI	Life Form	82.720	75.697	73.311	74.179	EcoBERT
WIKI	Life Form	81.360	71.838	70.318	70.653	DistilBERT
WIKI	Life Form	77.200	65.441	64.204	63.750	DeBERTav3
WIKI	Life Form	78.560	67.669	65.946	65.711	ELECTRA

Table 8: Model Results on the Categorical Trait Classification of Spanish and German Descriptions Task.

Dataset	**Trait**	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**Model**
WIKI_ES	Growth Form	70.733	49.172	21.899	28.335	Regex
WIKI_ES	Growth Form	80.700	75.653	76.073	75.856	Logistic Regression
WIKI_ES	Growth Form	85.056	82.049	80.782	81.364	DistilBERT
WIKI_ES	Growth Form	85.056	81.994	80.721	81.300	Multilingual DistilBERT
WIKI_ES	Growth Form	85.363	82.619	81.036	81.752	Spanish BERT
WIKI_ES	Growth Form	83.316	80.837	77.573	78.799	German BERT
WIKI_ES	Life Form	80.000	0.000	0.000	0.000	Regex
WIKI_ES	Life Form	67.391	62.265	61.074	61.532	Logistic Regression
WIKI_ES	Life Form	68.259	60.822	58.684	59.358	DistilBERT
WIKI_ES	Life Form	53.754	37.148	42.208	37.593	Multilingual DistilBERT
WIKI_ES	Life Form	74.915	67.840	67.432	67.516	Spanish BERT
WIKI_ES	Life Form	50.000	37.410	35.834	30.174	German BERT
WIKI_DE	Growth Form	74.125	81.863	42.566	47.967	Regex
WIKI_DE	Growth Form	85.026	79.890	78.940	79.390	Logistic Regression
WIKI_DE	Growth Form	86.356	82.074	82.773	82.343	DistilBERT
WIKI_DE	Growth Form	89.465	86.323	84.306	85.246	Multilingual DistilBERT
WIKI_DE	Growth Form	87.392	83.980	81.102	82.394	Spanish BERT
WIKI_DE	Growth Form	90.328	87.392	86.713	87.045	German BERT
WIKI_DE	Life Form	80.563	66.275	4.959	8.963	Regex
WIKI_DE	Life Form	74.178	69.276	70.102	69.528	Logistic Regression
WIKI_DE	Life Form	67.371	72.320	56.137	52.281	DistilBERT
WIKI_DE	Life Form	60.563	50.499	47.397	47.319	Multilingual DistilBERT
WIKI_DE	Life Form	69.014	62.197	62.194	61.438	Spanish BERT
WIKI_DE	Life Form	86.855	86.347	80.522	82.214	German BERT

Table 9: Model Results on the Categorical Trait Classification in a Data Deficient Regime Task.

Dataset	Trait	Training Set Sample Size	F1-Score	Model
POWO	Growth Form	32	27.744	DistilBERT
POWO	Growth Form	128	27.744	DistilBERT
POWO	Growth Form	512	79.685	DistilBERT
POWO	Growth Form	32	30.670	SetFit
POWO	Growth Form	128	79.468	SetFit
POWO	Growth Form	512	83.547	SetFit
WIKI	Growth Form	32	22.876	DistilBERT
WIKI	Growth Form	128	23.068	DistilBERT
WIKI	Growth Form	512	83.881	DistilBERT
WIKI	Growth Form	32	51.924	SetFit
WIKI	Growth Form	128	82.168	SetFit
WIKI	Growth Form	512	82.971	SetFit

Table 10: Model Results on the Numerical Trait Extraction Task.

Dataset	Trait	nMAE	Coverage	Model
POWO	Plant Height Max	7.559	26.917	DistilBERT
POWO	Leaf Length Max	30.537	27.826	DistilBERT
POWO	Leaf Width Max	23.157	60.965	DistilBERT
POWO	Plant Height Max	44.002	46.591	NumBERT
POWO	Leaf Length Max	39.983	43.132	NumBERT
POWO	Leaf Width Max	58.779	52.605	NumBERT