Contents

1 Introduction

The ever increasing urgency of climate change has resulted in enhanced regulatory requirements for institutional investors to incorporate environmental, social and governance (ESG) considerations into their portfolio decisions (Faccia et al., 2021). However, given the lack of alignment on how the companies should report their environmental and social impact or governance attributes (Jacobs and Levy, 2022), investors must rely on analysing vast quantities of unstructured, textual data to derive ESG scoring for their potential investments. Since manual scanning of text is a very time-consuming task that is prone to human error, investors increasingly turn to machine learning and more precisely, natural language processing (NLP) algorithms, which have the potential to classify unstructured textual data into meaningful ESG score categories(Sokolov et al., 2021).

This trend has emerged as a result of the recent introduction of transformer-based models such as GPT-3 (Radford et al., 2018) or BERT (Devlin et al., 2019) that brought considerable gains in predictive accuracy for natural language understanding and comprehension tasks. Particularly, these models have shown to perform especially well when pre-trained on a large corpus of data and subsequently, fine-tuned on a downstream classification task (Wolf et al., 2020).

Consequently, recent years have seen increased utilisation of these models in business applications as a means to replace or improve upon their prior processes (Bahja, 2021). Sokolov et al., 2021, utilise a BERT classifier, fine-tuned on a large corpus of climate related news data to successfully incorporate information on potential ESG issues into a portfolio management model, producing improved results on three of four key indicators for well performing portfolios.

Furthermore, Antoncic et al., 2020, utilised several different NLP and machine learning models to analyse company ESG reports, aiming to remove the resulting self-reporting biases and thereby more correctly estimate the true ESG scores for each company.

While these applications show some considerable success in the analysis of climate related corpus and the subsequent integration of their results in processes such as financial management and risk scoring, several shortcomings of the early stage transformer models remain. In particular, the generic nature of the train dataset of models such as BERT often result in poor understanding of corpuses with topic specific language such as ESG goals (Lee et al., 2019; Beltagy et al., 2019). Hence, large amounts of data are needed during fine-tuning to correctly capture and understand the specific nuances in the meanings of these texts.

Unfortunately, whilst a vast amount of unlabeled text instances can be collected at once, collection of labelled examples has continuously shown to be an expensive and time-consuming process (Grießhaber et al., 2020) and thus, the wide-spread adoption of effective ESG scoring classifiers for business purposes has not yet occurred. The ESG case serves as the perfect example case to describe two of the largest bottlenecks of the widespread adaption of NLP models in a business setting: low resource scenario, in particular shortage of labelled data and domain specificity. In fact, since the introduction of the transformer various other research has attempted to address the issue by increasing parameter efficiency of the fine-tuned model or by maximising the information gain that can be obtained from a small set of examples. (Houlsby et al., 2019; Grießhaber et al., 2020)

One such research, aimed at tackling the amount of trainable parameters in a transformer-based model that arise due to the model's inherently deep structure, Houlsby et al., 2019 proposed adapter modules, which rather than fine-tuning the full model, train only the newly introduced set of weights at every layer. Such a fine-tuning strategy can result in elimination of up to 99.9% trainable parameters, while achieving comparable accuracy (Houlsby et al., 2019). As a consequence, the adapters prove more robust to overfitting and demonstrate better generalisation ability, which in turn results in improved performance in low-resource settings with domain-specific tasks (Houlsby et al., 2019).

On the other hand, a commonly deployed paradigm developed to cope with limited annotation resources is active learning (Settles, 2010). The approach tries to alleviate the labelling burden by identifying the most informative data instances as these maximise the knowledge gain that the model can attain when training on the respective data (Burr, 2009). Some initial studies have focused on exploring the suitability of this approach when applied to deep pre-trained models that were subsequently fine-tuned on low-resource data, yielding promising results (Grießhaber et al., 2020). Nonetheless, since the fine-tuning is performed on all weights of the base model (Grießhaber et al., 2020), the previous active learning experiments can be considered as highly parameter inefficient and thus, can result in overfitting when only a small train sample is available.

All in all, given the proven robustness of adapters when deployed to domain-specific tasks coupled with the potential of active learning to uncover the most informative labelled instances in low-resource scenarios give rise to a query whether tasks suffering from this resource constraint, such as ESG assessments, could benefit from deploying active learning approaches to adapter-base tuning. If successful, the introduction of such an algorithm could significantly impact the adoption of NLP models for business applications, particularly within smaller firms who typically have less data available.

To the best of our knowledge, there has been no previous research that would systematically explore the performance capabilities of applying the active learning algorithm on an adapter-based model in the fine-tuning stage of transfer learning. Thus, this gives a rise to the research question:

Does deploying active learning to fine-tuning adapter-based classifiers improve performance on tasks with resource constraints characterised by thematic language and small sample size?

In order to investigate this issue, we develop a novel algorithm, P-BAAL: Pool-Based Adapter Active Learning, which intelligently combines the techniques of the two above described approaches. Additionally, we opt to utilise an unstructured dataset consisting of

national adaptation and mitigation plans provided by countries that pledge their commitment to the Paris Agreement. The data is currently exploited by an institutional investor, Climate Fund Managers B. V., who classifies whether the documents are relevant in affecting ESG impact scoring. This peculiar choice of dataset ensures a close alignment of the theoretical model outlined in our research paper and its performance with respect to an applied scenario such as ESG assessments.

This paper will be structured as follows. In Section 2, a detailed overview of several recent state-of-the-art NLP models and further research into their adaptability to constrained environments is provided. Subsequently, we develop a set of hypotheses to investigate the research question in Section 3. Building on the research hypothesis, we design the model architecture in Section 4. Furthermore, in Section 5 we provide a brief discussion and preliminary analysis of the dataset as well as any preprocessing steps taken, whereas the P-BAALI algorithm, as well as base models, adapters and hyperparameters utilised, are outlined in Section 6. Consequently, we provide an overview of the results in Section 7, followed by a discussion of these with respect to the research question in Section 8. Finally, we finish the paper with a conclusion, containing any limitations to our approach and suggestions for potential further research.

2 Contributions

1. We propose a novel mechanism that applies pool based active learning to adapter-based models and delivers state-of-the-art performance when applied to classification tasks in the field of natural language inference characterised by low data availability and domain specific language

2. We show that additional performance improvements of P-BAAL are conditional to an appropriate definition of the increment at which new labelled instances are included to the model during active learning iterations. In detail, the model seems to perform

specifically well when the data increment is defined as an exponentially decreasing function as the ratio of the initially available sampling pool.

3. We demonstrate the applicability of the proposed theoretical model to an applied business scenario by utilising P-BAAL to train a classification model, which, given its superior performance, has been adopted by an institutional investor to automate its ESG assessment of potential investors.

3 Related Work

This section will be structured as follows. Firstly, an overview of several different transformer architectures used in the subsequent analysis will be provided, including any shortcomings for each of these. We then outline the general problem of such base models to provide high accuracy in situations such as the ESG issue described in Section 1, which are characterised by resource constraints and highly thematic language. Consequently, an overview of the two fields of study whose combination serve as the base for this research question is provided, with a detailed description of their efforts to alleviate the constraints at hand.

3.1 Language models as a successful use case for natural language comprehension

When it comes to the tasks of language understanding and natural language generation, recurrent neural networks (Hopfield, 1982) and long-term short-term memory networks (Hochreiter and Schmidhuber, 1997) had long been considered as the state of the art approaches (Vaswani et al., 2017). However, due to their inherently sequential nature, which prevents parallelization, these models suffer from computational inefficiencies and memory constraint especially when dealing with large datasets or long sequence lengths

(Vaswani et al., 2017). To address these issues, Vaswani et al., 2017, proposed an alternative model architecture, the Transformer, which circumvents the recurrence by relying solely on attention mechanisms to derive dependencies between inputs and outputs. As a result, not only the model's scalability in model parameters and training data increased drastically, but also considerable gains in predictive accuracy were achieved when the model is pre-trained on a large corpus of data and subsequently, fine-tuned on a downstream task (Wolf et al., 2020). However, constraining the model to deploying unidirectional self-attention layers to determine word dependencies restricts the architecture that can be deployed during pre-training, which in turn considerably limits the model's performance during the fine-tuning for sentence level tasks (Devlin et al., 2019). This is due to the fact that for token based tasks such as question answering, it is essential to incorporate information flow from both direction to derive context and thus, it is suboptimal to be dealing with attention layers, whose tokens can only attend to their previous tokens (Devlin et al., 2019).

Inspired by Taylor, 1953, Devlin et al., 2019, propose an extension of the transformer model, Bidirectional Encoder Representations from Transformer (BERT), to alleviate this constraint. BERT achieves a bidirectional flow of information by exploiting a masked language model (MLM) during pre-training to obtain word representations. In particular, MLM randomly masks few of the input tokens and thus, enables the fusion of the right and left context (Devlin et al., 2019). At the same time, to improve reasoning about relationships between pairs of sentences, the authors incorporated next-sentence prediction, which tries to determine whether two passages follow each other in the original text (Devlin et al., 2019). As a result, the model brings advances in the ever-achieved accuracy for eleven natural language processing domains, overperforming complex, heavily-engineered task-specific architectures for both sentence level and token level tasks (Devlin et al., 2019).

It had soon become apparent however, that the model's capacity was not fully exploited as subsequent research started pointing to the fact that the underlying architecture was

still substantially underfitted. Furthermore, only a limited combination of hyperparameters was trialled out during the training procedure (Liu et al., 2019), potentially resulting in a suboptimally trained model. Therefore, Liu et al., 2019, focused on identifying the appropriate design choices and training strategies and as a result proposed RoBERTa, a robustly optimised BERT pre-training approach, which successfully delivered improvements on downstream task performance. Among the tactics found as most effective for performance enhancement, the authors outline the inclusion of longer training periods, with bigger batch sizes and more data, which should be concurrently characterised by longer underlying sentences. Additionally, dynamic, rather than static, masking was introduced, which allowed for the generation of a new masking pattern every time a specific sequence was fed into the model (Liu et al., 2019). Lastly, whilst Devlin et al., 2019 found the next sentence prediction to be a crucial training component for context comprehension, Liu et al., 2019, showed that if the model is fed full sentences rather than sentence pairs as input during pre-training, removing next sentence prediction matches or enhances performance during fine-tuning.

With all the aforementioned progress and the promising initial results outlined above, the models' pre-trained architectures started to be dominantly exploited for several transfer-learning tasks (Sanh et al., 2019). However, as contemporaneous research, which focused on the optimization of pre-trained models, pointed towards a close, positive relationship between the size of a model and its performance on downstream tasks, the wide adoption of ever more computationally and memory demanding models started to be questioned (Sanh et al., 2019). To alleviate these inefficiencies, Sanh et al., 2019, utilised knowledge distillation, in which a "student" model tries to replicate the behaviour of the "teacher model", to develop a much smaller counterpart of the BERT model, DistilBERT, which required a considerably smaller training budget and was characterised by faster inference time, whilst preserving the original model's predictive performance. Using a similar reasoning, a distilled version of ROBERTa was subsequently introduced (Sanh

et al., 2019).

Nevertheless, such distillation techniques impose considerable restrictions on the resulting model as each layer of the "student" model is required to follow the same structure as that of its "teacher" (Wang et al., 20200). Thus, rather than performing per-layer distillation, Wang et al., 2020, proposed to focus only on the self-attention module of the last layer found in the "teacher's model", to address the difficulties in layer mapping between the two models and hence, to enable more flexibility for the layer design of the respective "student" model. The newly developed MiniLM, characterised by various parameter sizes, succeeded in outperforming the baseline models (Wang et al., 2020).

As both BERT and RoBERTa and their distilled counterparts expanded the boundaries of what was thought of as state-of-the-art results for many natural language processing tasks, subsequent research focused on refining the models' design to develop domain-specific architectures that, in-turn pushed those boundaries even further. Given that the research topic of this paper is closely related to natural language comprehension and information retrieval, we next focus on describing a group of architectures that were developed specifically for this domain.

When it comes to comprehension tasks such as large-scale similarity comparison or information retrieval, the aforementioned models were for a long time considered inapplicable (Reimers and Gurevych, 2019). This was due to the large computational overhead that arose as a result of tedious inference computations, which had to be performed to identify similar pairs in the underlying corpus (Reimers and Gurevych, 2019). To address these inefficiencies, subsequent research focused on deriving fixed size sentence embeddings by exploiting the BERT output layer or by simply utilising the output of the first, [CSL] token (Reimers and Gurevych, 2019). However, the results achieved were rather suboptimal, underperforming even the GloVe embeddings (Pennington et al., 2014).

Therefore, Reimers and Gurevych, 2019, proposed BERT and RoBERTa extensions, Sentence BERT (SBERT) and Sentence RoBERTa (SRoBERTa), which proved success-

ful for the extraction of meaningful sentence embeddings via the addition of a pooling operation to the output of BERT and RoBERTa, as well as the utilisation of siamese and triplet networks to update the weights during fine-tuning. As a result, semantically related sentences can be promptly uncovered in large data corpuses, by performing a euclidean or manhattan distance calculation on the retrieved embeddings (Reimers and Gurevych, 2019). At the same time, the new architectures archive state-of-the art results for natural language inference tasks (Reimers and Gurevych, 2019).

3.2 Limitations of standard fine-tuning on general pre-trained models

Even though all the above models are capable of delivering satisfactory results when fine-tuned on the downstream tasks, the transfer learning phase is quite parameter inefficient since all the pre-trained weights of the underlying model have to be retrained (Houlsby et al., 2019). Whilst this would not necessarily cause a problem when the fine-tuning is performed on large corpus of data, labels are typically costly to obtain and unfeasible on a large scale for the majority of organizations operating within specialised fields such as environmental impact investing and ESG issues (Nugent et al., 2020).

In particular, utilising transfer learning when only a small set of labelled examples is available can lead to overfitting as the model learns complex, yet incorrect relationships between inputs and outputs due to the sampling noise present in the training set (Nugent et al., 2020). Additionally, when utilising a base model, which was previously trained on general corpus, for fine-tuning purposes on a downstream task, which is characterised by a domain specific language, the base model will not be able to adjust its pre-trained weight to an extent such that it would learn to recognize the thematic nuances of the new text corpora (Nugent et al., 2020; Beltagy et al., 2019).

As the application of a model for the purpose of ongoing business operations, such as

scoring investment projects on their environmental impact, requires at least near-human performance, these roadblocks make it oftentimes infeasible to apply a generically fine-tuned model for such a process. Thus, in scenarios when new training data is difficult or time-consuming to obtain, addressing parameter inefficiency of neural networks can be crucial to alleviate the aforementioned problems, which would allow for the development a model that can be utilised to fully automate a business process.

3.3 Exploiting adapters to improve the parameter efficiency in fine-tuning procedures

One of the most prominent alternatives to transfer learning that addresses parameter inefficiency can be considered an adapter, originally proposed by Houlsby et al., 2019. Adapters are compact neural networks characterised by a small number of parameters and near-identity initialization, which are inserted in between or in parallel to layers of the original model. Unlike in classic transfer learning, where the original weights and the newly introduced classification layer are trained jointly, in adapter tuning the parameters of the original model remain untouched and only the adapter modules are trained (Houlsby et al., 2019).

The originally developed adapter design, the bottleneck adapter, projects the input into a lower dimensional space, and subsequently applies a nonlinear activation function before it maps the input back into the original space and lastly, adds a residual connection (Houlsby et al., 2019). These architectures are appended to each transformer sub-layer after both the multi-head attention and feed-forward block (Houlsby et al., 2019). As the authors find, such a design enables the elimination of 96.4% of trainable parameters and delivers comparable results to when a whole base model is fine-tuned.

Inspired by its initial success, the interest of academia quickly shifted towards experimenting with different adapter architectures, in order to discover effective designs that

would deliver enhanced performance, while further diminishing the number of trainable parameters (He et al., 2021). One of the initial improvements, the prefix-tuning adapter, was developed by Li and Liang, 2021. In their approach, the authors only prepend a tunable vector to the pre-trained keys and values found in the multihead attention blocks of the base model (Li and Liang, 2021) , as opposed to the bottleneck adapter, which inject the modules after both the feed forward sublayers and the attention blocks (Houslby et al., 2019). As a result, the prefix-tuning adapter utilises a mere 0.1% of parameters of the original model, and outperforms not only the base model for fine-tuning in low-data settings (Li and Liang, 2021), but also the bottleneck adapter on a multitude of tasks (He et al., 2021).

As the adapter configurations outlined above have been originally proposed as standalone, He et al., 2021, decided to investigate whether it could be beneficial to combine the adapters in order to exploit strengths in their respective designs. In their study, the authors proposes a Mix and Match adapter (MAM), which utilises an efficient combination of prefix tuning introduced by Li and Liang, 2021 and parallel bottleneck adapters, whose activations are passed in parallel to the adapted sub-layer of the base model rather than in sequential order (He et al., 2021). In detail, the authors try to minimise the number of trainable parameters of the prefix tuning to control for parameter efficiency, while allocating a larger parameter budget for the parallel adapter to enhance the performance on the downstream task (He et al., 2021). As a result, the MAM adapter outperforms previous adapter design configurations, while training 93.3% less parameters than in the transformer model.

Eventhough the adapters have been originally proposed to address parameter inefficiency in fine-tuning, subsequent studies showed that the unique characteristics of adapter-tuning that allows for alternation between frozen and learnable layers also leads to better generalisation ability (He et al., 2021). Consequently, the approach yields superior results in low-resource settings, especially when the data to train on is domain specific

(He et al., 2021).

3.4 Active learning

The introduction of adapters brought dramatic parameter efficiency benefits coupled with performance enhancements especially considering settings when only a small sample of data is available for fine-tuning purposes (Li and Liang, 2021). Nevertheless, it would be unreasonable to expect that relying solely on diminishing the number of trainable parameters can eliminate all sampling noise that is generally present in small datasets. Thus, to further alleviate this sampling bias, the fine-tuned model must by some means develop the capacity to maximise the knowledge gain it can obtain from the train sample. More precisely, it must be capable of identifying data instances that carry unique and valuable information, which can be exploited to enhance accuracy.

A similar theory motivates active learning, which is a subfield of artificial intelligence closely related to machine learning, whose principal idea consists of allowing the underlying algorithm to be 'curious' during the training phase, meaning the algorithm is free to choose the data that it will learn from (Burr, 2009). Subsequently, the chosen labels are presented to an oracle, such as a human annotator, who provides the model with the truth and the model alters its weights accordingly (Burr, 2009). Whilst a multitude of strategies for identification of the relevant train instances were developed, this section further focuses on describing pool based active learning as its data selection technique closely relates to the idea of information gain maximisation (Burr, 2009).

In pool based learning, the model M with parameters θ gradually selects q number of instances X_i from the pool U according to an acquisition function $a(x, M)$, which tries to maximise possible knowledge gain of the model when trained on X (Grießhaber et al., 2020). After being labelled by the oracle, the chosen data points are added into the training dataset $train$, according to the logic outlined in Equation 1 (Grießhaber et al., 2020).

$$train_{new} = train_{old} \cup U_{x \in U_{argmax}(a(x,M))} \tag{1}$$

Even though multiple approaches to defining $a(x, M)$ have been proposed, we further focus on those that showed promising results in early research on active learning applied to information extractions tasks (Culotta and McCallum, 2005; Settles, 2010). One of the most widely utilised strategies can be considered uncertainty sampling, due to its strong performance and straightforward adoption for probabilistic learning models, such as for the model's classification head trained during fine-tuning (Burr, 2009). In detail, after multiple forward passes are performed on the same input, the instances for which the classification results produced the most disagreeing predictions with high uncertainty - highly informative - can be identified by analysing the normalised output layer of the model, (Grießhaber et al., 2020). Thus, these most informative instances will be identified by maximising the entropy of the predictions such as defined in Equation 2 (Shannon, 1948).

$$X^*_{ent} = \arg\max_x - \sum_i P(y_i|x;\theta) \log P(y_i|x;\theta) \tag{2}$$

The maximisation of approximate knowledge gain of the model, called uncertainty sampling, has the potential to ameliorate performance in low-resource scenarios since the model is inherently programmed to choose instances that capture diverse semantic aspects and thus, the risk of overfitting commonly present when trained on little data can be decreased (Bashar and Nayak, 2021).

The uncertainty sampling algorithm leverages the predictive power of only one adapter based model at a time, which might not be capable of capturing different regions of the version space (Burr, 2009). To alleviate the issue, a query by committee (QBC) active learning approach had been proposed, which trains a multitude of "committee" models $C = \{\theta^{(1)}, \ldots, \theta^{(C)}\}$, and subsequently, rather than including new instances according to

per-model uncertainty, the approach selects those data points, for which the underlying models produced the most disagreeing predictions (Burr, 2009). Therefore, the information gain can be maximised by selecting instances with the highest variance in predictions (Burr, 2009). Even though there is no upper limit on the number of models to be included in this type of active learning, interestingly, query by committee has shown to work well with small committee size, which included only 2 or 3 models (Burr, 2009).

4 Methodology

As introduced in Section 1, in this study we try to ameliorate the low performance capabilities of pre-trained transformer-based models when fine-tuned for information retrieval tasks using data, which is not only limited, but also possibly domain specific. Since previous research successfully shown that training the parameters of adapter modules rather than pre-trained weights of the transformer leads to a surge in the posterior predictive accuracy in low-resource scenarios (Li and Liang, 2021), we opt to perform all fine-tuning runs exclusively on the adapters, following the training approach proposed by Houlsby et al., 2019.

Nevertheless, as outlined in Section 3, it is improbable that shrinking the amount of trainable parameters can fully address the sampling noise generally present in small datasets. Thus, it is crucial to develop a fine-tuning strategy that would enable for identification of training instances that can be considered as most informative for training purposes. Since the underlying principle of active learning is to let the model freely choose the labels to train on based on some informativeness measure Grießhaber et al., 2020, in this study we investigate whether incorporation of such approaches to adapter-transformer models could bring the desired performance advancements. Few recent studies have already explored various deep-active learning approaches to text classification tasks in low-resource scenarios (Ein-Dor et al., 2020, Grießhaber et al., 2020, Schröder et al.,

2021, Ikeagami et al., 2022), all of which delivered promising results. Nevertheless, these studies all share a common drawback, that is the requirement of fine-tuning all, or at least a big proportion, of the pre-trained weights w of the base model, which is suboptimal in situations of small samples with domain specific language, as it results in the tendency to overfit. Therefore, it can be expected that leveraging a technique that trains a substantially smaller subset of parameters, could bring advancements to the results achieved so far. Opportunely, the adapter modules were developed with a specific purpose to address parameter inefficiency when fine-tuning transformer base models. Whilst, no contemporary research has thus far focused on the evaluation of the combined predictive power when active learning is exploited on adapter-based models, their joint utilisation has the potential to deliver promising results. This leads to the development of the following hypothesis:

Hypothesis 1: *Applying pool-based active learning with uncertainty sampling when performing transfer learning on the weights of the adapter layers can lead to increase in posterior predictive accuracy on the downstream information retrieval task, in scenarios with little labelled data available irrespectively of the adapter structure chosen for training.*

As stated in the hypothesis, the resulting training method, Pool-based Adapter Active Learning (P-BAAL), applies the common active learning approach of uncertainty sampling to an adapter model to provide it with only the most informative pieces of data in the fine-tuning stage.

When applying P-BAAL, it is reasonable to expect that the increment by which the training data grows at every training re-initiation has a diminishing effect on the benefits introduced by active learning. This is due to the fact that active learning assumes the model is programmed such that it is capable of improving its performance through the identification of instances that bring the largest knowledge gain. If we force the model to incorporate too much additional data with little useful or repetitive information, the model loses the ability to recognize and eliminate sampling noise of the data and as a result, the

risk of over-fitting becomes more prevalent. This leads to the development of Hypothesis 2, which states:

Hypothesis 2: *There exists a negative relationship between the data increment at each training iteration and the performance of Pool-based Active Adapter Learning (P-BAAL) .*

While the data increment at each iteration in the uncertainty sampling method used throughout Hypotheses 1 and 2 assumes a constant increase, it could be reasonable to assume that a non-constant increase in the additional sample size would provide a boost to performance. Specifically, the incorporation of a larger proportion of new instances for early training iterations would help with faster convergences and prevent overfitting, by incorporating larger amounts of informative data early on. At the same time, once the pool of data excluded from training becomes sufficiently small, the model should add only a few new instances for each subsequent iteration, since it can be assumed that the most informative data has been already identified. Thus, the few additionally added inputs should only eliminate any remaining small imprecisions. Such dynamics for new instances incorporation can be achieved by specifying the number of additional samples at each iteration as a function of the initial training size and is defined in more detail in Section 6 . Consequently, the third hypothesis can be formulated as follows:

Hypothesis 3: *Modelling the increment in the additional training sample as a declining function of the number of iterations has a positive effect on the performance of P-BAAL as compared to a constant sample increase.*

As discussed in Section 1, we aim to directly tie the research carried out in this paper to real world business applications, in particular the utilisation of NLP models to improve assessments on ESG reporting. Hence, the dataset introduced in detail in Section 4 contains a set of document pairs of investment proposals and their potential ESG relevance match, sourced from an organisation that utilised this data on a daily basis. The choice

of dataset allows us to not only gain an understanding on the general ability of our approach to improve performance on classification tasks, but also get insights on their direct applicability in a production setting.

5 Dataset

This section provides a detailed overview of the dataset originally created by the author of this paper for the purpose of ESG relevance understanding for CFM and subsequently restructured for this research. Due to the complexity and unique design of the dataset, we decided to firstly provide a detailed explanation of the unstructured data and the various processing steps taken to arrive at a dataset following the classic premise-hypothesis setting of natural language inference (Jiang and de Marneffe, 2019). The resulting dataset contains a set of unlabelled instances 37,456 hypotheses pairable with the potential set of premises consisting of all potential combinations of a set of 97 themes.

This complete set is subsequently analysed to provide a complete overview of the data distribution irrespective of whether it is used for training purposes. Furthermore, we outline the steps taken to derive and correctly label the subset of samples used for model training, evaluation and testing, before finally providing insights on any difference between the training set and the complete corpus in the initial analysis.

5.1 Origin of source data: addressing business process inefficiencies through the utilisation of NLP

Climate Fund Managers B.V. (CFM) is an impact investment manager, financing projects that are operating in climate themed areas. Since CFM mobilises its capital from both private and public sectors, its underlying investment vehicles must account for a wide range of funding disbursement constraints imposed by its investors. One of the restrictions re-

lates to the mandatory reporting on the Rio marker score (OECD Development Assistance Committee, 2022) for each potential investment considered for inclusion to the investment pipeline. This score is determined based on how closely the investment project addresses the 4 themes of Rio Conventions, namely biodiversity, desertification, climate change mitigation and climate change adaptation (OECD Development Assistance Committee, 2022), while accounting for the specific needs of a country where the project company will be operating.

To facilitate the assessment and to assure reporting accuracy, the OECD Development Assistance Committee defines a set of publicly accessible documents that can be used for score justification. These are periodically issued by governments around the world and fall into one of the following categories- the National Determined Contributions (NDCs), National Adaptation Plans (NAPs) and National Adaptation Programmes of Actions (NAPAs). Even though all these documents carry climate change mitigation and adaptation as their central topic, the NDCs put focus on outlining the definite action each country commits to take in order to achieve the goals of the Paris agreement (United Nations for Climate Change, n.d.-b), whilst NAPs and NAPAs list activities specifically tailored to the needs of grassroot communities that enhance countries' ability to adapt to adverse effects of climate crisis (United Nations for Climate Change, n.d.-a,United Nations for Climate Change, 2020).

Originally, CFM relied on human capital to manually scan the above-mentioned documents for relevant paragraphs that could justify a close alignment of an investment proposal with the climate change mitigation and adaptation plans outlined as crucial for the country of operation of the potential investee. However, such an approach is not only time inefficient, but also error prone as the documents are quite lengthy and technical and thus, the relevant passages were many times overlooked.

To address these reporting inefficiencies, CFM had decided to explore the possible application of natural language understanding as the task at hand can be partially formulated

as a semantic similarity detection problem. More precisely, the core idea was to determine the relevance of an investment proposal (premise) to passages found in NDCs, NAPs and NAPAs (hypothesis).

5.2 Initial data collection

The investment proposals that are collected by CFM come in various formats, their structure and content are very diverse and possibly carry implicit bias (e.g., favouritism) introduced by the author of the investment pitch (Antoncic et al., 2020). Therefore, rather than utilising the full investment proposal as premise, we opted to categorise the documents to a large set of predefined topics that had to satisfy two conditions. Firstly, the topics had to be closely aligned with the themes determined in Rio Marker Conventions and secondly, these themes had to fall within the investment scope of the fund vehicle. In detail, the propositions had to not only relate to biodiversity, desertification, climate change adaptation or mitigation but also fall within the area of water and waste-water infrastructure.

Utilising this approach, we have collected a final list of $N = 97$ descriptions of climate themes, serving as the set of unique topics whose potential combinations represent the entire set of potentially definable premises. More precisely, the set of potential premises C is described in Equation 3 and can be described as the set of all k-combinations, for any positive k less than or equal to 97,: of the set $S = \{A_1, \ldots, A_n : N = 97\}$ consisting of N climate topic descriptions A.

$$C = \bigcup_{k=1}^{N} S_k^N \tag{3}$$

Such data design assures that the assumption of train and test data being drawn from the same target distribution will be more likely to be satisfied as the context and the underlying word distribution of premises will remain fairly constant over time. This provides a particular applicability in such a production setting as the ESG relevance assessment

of each incoming investment proposal can be reduced to an assessment of the relevance of the combination of themes attributed to said proposal. Additionally, we have collected a full collection of documents outlined by OECD Development Assistance Committee that were subsequently used as the set of hypotheses.

So far 193 parties have issued NDCs, 81 countries have developed NAPs and 51 have submitted NAPAs. However, since the investment scope of CFM is strictly focused on developing countries across Africa, Asia and Latin America, not all of the documents were relevant for the analysis and were thus discarded to try to minimise the amount of manual data cleaning. Since NAPs and NAPAs are generally issued by lower-income economies, the data loss affected the NDC type of documents more severely. Nevertheless, given that NDCs are on average shorter than NAPs and NAPAs, an average of 5 to 10 pages per document compared to an average of 20 to 100 pages per document respectively, the amount of data omitted from the analysis can be considered minimal.

Since CFM receives all of the above documents in pdf format, an Optical Character Recognition algorithm had to be deployed to transform the information into an appropriate format usable for training purposes. Subsequently, to assure high data quality, we performed manual cleaning on the transformed text. More precisely, we re-written sections that were not transformed correctly by the algorithm and additionally, we eliminated any redundant passages from the text, such as tables of content, section titles or page numbering. Finally, to assure that the algorithm can correctly identify relevant hypotheses for the individual premises, the preprocessed texts were splitted from documents into paragraphs, which can be considered structural units of text consisting of sentences relating to one main idea (Marsh, 1984). Leveraging such an approach, we arrived at a dataset of 37,456 potential hypotheses each of which could be relevant to any of the 97 themes whose combination is defined as premise. It should be noted, that due to the static nature of the hypotheses dataset, whose documents are only sparsely updated on a periodic basis, and the pre-definition of the set of potential premises via the 97 topics, the complete

set of potential input data viewable in a production setting is actually available in unlabeled format.

Thus, we will now dive into an analysis of the complete dataset to provide insights on the true distribution of some key indicators such as paragraph length or term frequencies, later usable for a more structured definition of the hyperparameters of the model. Furthermore, it will allow us to validate the alignment between the distribution of the final labelled test set and the complete dataset at hand.

5.3 Data analysis of the complete dataset

Having collected the final databases of topics and of the supplementary documents, we performed basic data analysis in order to gain better understanding on whether there are some data-specific features that need to be taken into account for the later definition of model parameters.

5.3.1 Climate topic themes

Firstly, we put under investigation the set of premises via an analysis of its underlying set of climate topic themes, which were previously developed internally at CFM. As shown in Figure 1, a single topic description may contain generally very few words up to about one sentence. This is expected as the topics are predefined to be short definitions of the main point in an investment proposal.

Additionally, we applied the term frequency-inverse document frequency (TF-IDF) measure to estimate how relevant a word in a collection of topics is, which when plotted against its inverse document frequency (IDF) calculates how common or rare a given word is (Jones, 1972). Thus, it can uncover the most commonly occurring important words across all topics. The resulting graph is shown in Figure 2. It can be concluded that the themes generally revolve around water use and supply for agricultural or afforestation purposes

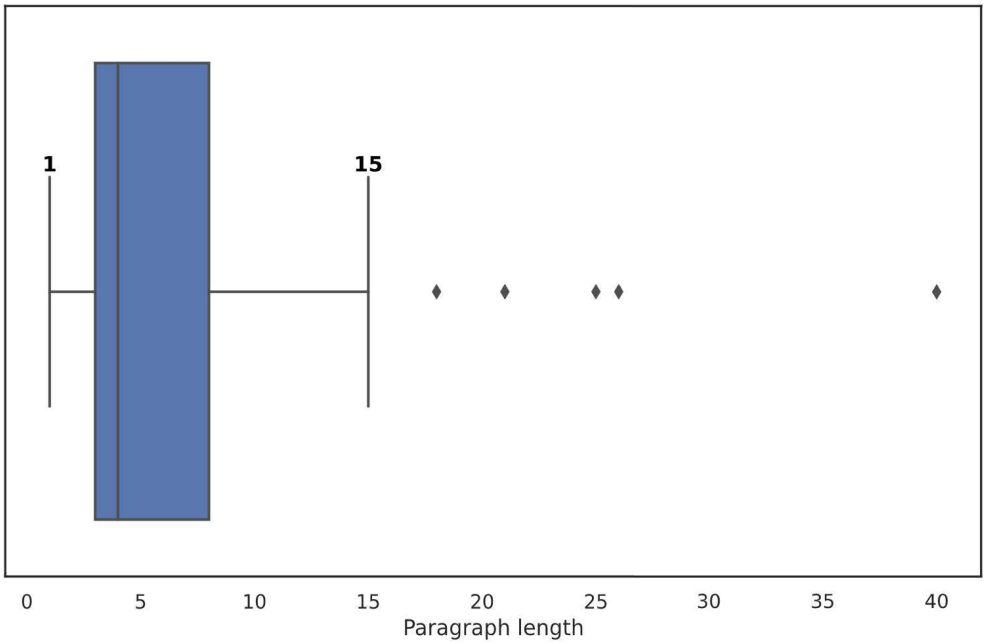


Figure 1: Boxplot of word count distributions for premises

as well as around wastewater management, which is closely aligned with the investment focus of CFMs fund vehicle.

5.3.2 Hypothesis dataset

Subsequently, we analysed the overall distribution of hypotheses across each of the three original document types. As can be observed in Figure 3, whilst NDCs are overrepresented in their document count, their content is considerably smaller resulting in a fairly even distribution across the three document types at 14,702, 12,247 and 10,507 for NAPAs, NDCs and NAPs respectively.

Furthermore, it is essential to understand the word count distribution in the paragraphs collected for the appropriate model configuration. It can be inferred from the Figure 4 that the distribution is skewed considerably to the left, with the values of each of the quartiles equal to 14, 27 and 55 respectively, while the outlier threshold can be set at 116 words per paragraph.

Additionally, to estimate the most prevalent themes within the hypotheses, we again

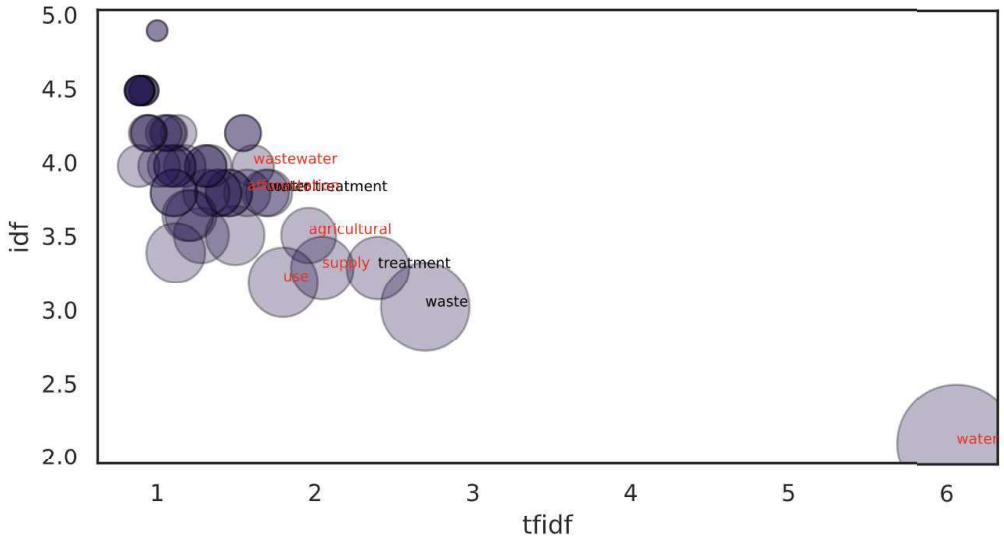
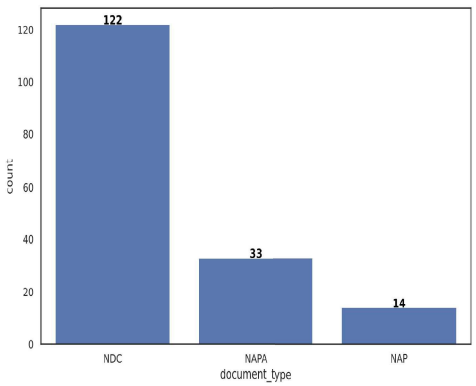
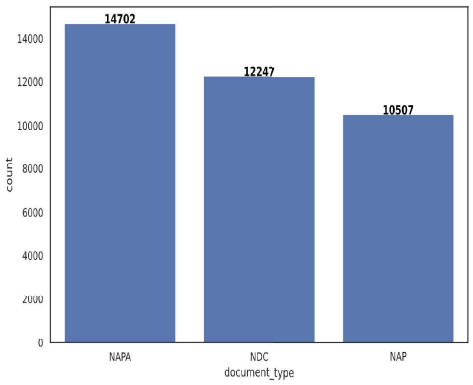


Figure 2: Most frequent words in premise dataset



(a) Unique documents per type



(b) Hypotheses across document type

Figure 3: Data distribution across document type

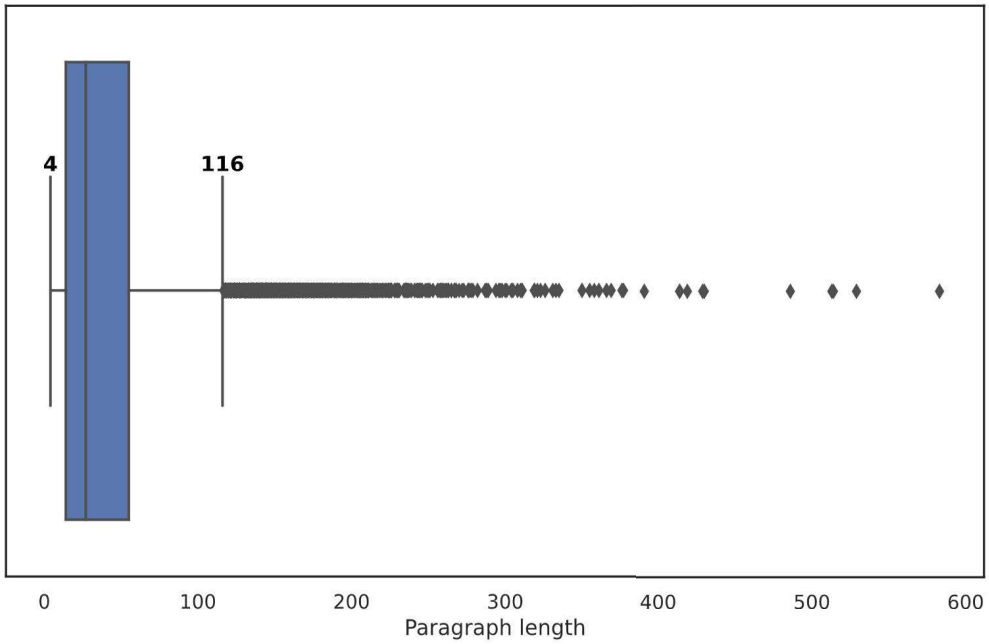


Figure 4: Boxplot of word count distributions for hypothesis dataset

estimated the TF-IDF measure for each word, which we subsequently plotted against its IDF. It can be inferred from Figure 5 that unlike the premise set, which is quite topic specific, the hypothesis set contains information more generally relating to climate change, resource management, habitat protection or adaptation practices. Furthermore, when we applied the same exercise on subsections of corpus each relating to only one of the three document types, we observed very little variation in the results, confirming the assumption that the themes of the NDCs, NAPs and NAPAs are closely aligned. For the graphical representations per document type, please see Figures 8, 9 and 10 in the Appendix.

5.4 Final dataset labelling and preparation

In order to utilise the data for this research, the previously collected and pre-processed data has to be further restructured into a format that can be fed into an algorithm for training. Therefore, it was necessary for us to sample a subset, that could be labeled, of the complete set of premises and hypotheses described in Section 5.2. As such we utilised a sample of premises based on previous project descriptions received by CFM for

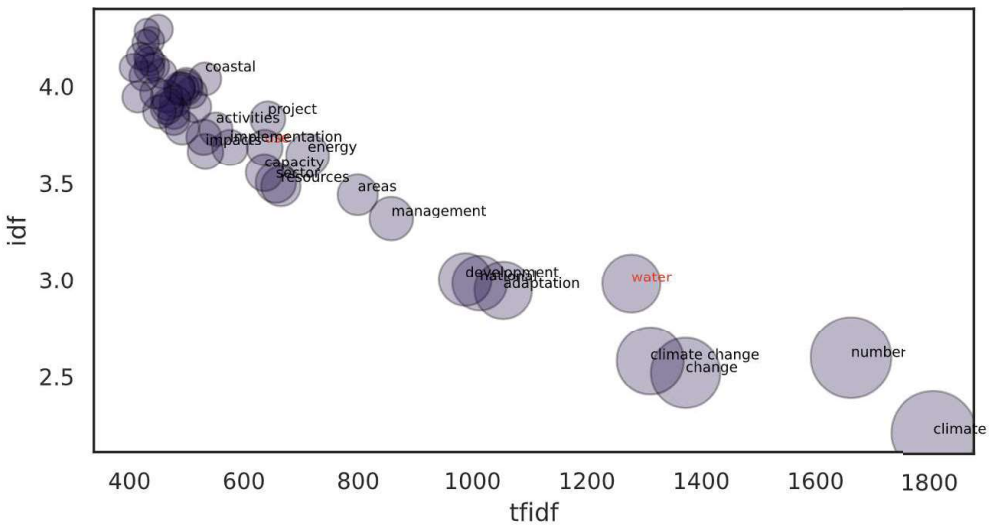


Figure 5: Most frequent words in hypothesis dataset

funding proposals. Furthermore any unique topic not considered in the previous set was included as an additional premise, amounting to a total of 120 unique premises. At the same time, we assigned to them examples from the hypothesis set that had been used for the score justification in early stages of CFM's operation.

Subsequently, for each pair in the dataset we manually assigned an output label such that it can be utilised for training purposes. As output classes, a simple binary classifier is utilised to highlight whether a given hypothesis is relevant to the premise. We decided on this as the task at hand does not actually require a deep understanding of the underlying sentiment, but simply whether any relation between two passages can be identified. On top of that, having a smaller number of output classes should in theory increase the model's performance, especially in low resource scenarios, when we have little training data available to start with.

5.5 Analysis of the final dataset

Utilising such an approach, we have arrived at a final dataset of 5321 of records that have been subsequently divided into training, evaluation and test sets according to a split ratio of 60/20/20 (3193,1064,1064). It should be noted however, that during the deployment of

active learning approaches the training commences with a subsample of the training set defined based on a ratio, which is tuned to achieve the optimal performance. For further details and results attained, please see Section 6 and Section 7 respectively.

As the underlying distribution of the topics from which the premises are derived are constant and thus will not change with respect to the training sample, any further analysis on the selected subset will focus on the distribution of the hypothesis set, to verify its representativeness of the complete set of hypotheses. As can be inferred from Figure 6, the average length of the subset hypotheses are significantly smaller when compared to the overall set collected from all documents outlined by the OECD Development Assistance Committee. Our initial analysis showed that this discrepancy can be attributed to the fact that generally the long paragraphs provide background on the country and its geographical and economic situation, whilst the shorter passages focus on outlining the actions to be taken with respect to climate adaptation and mitigation. Additionally, given that the outlier threshold for the paragraph distribution lays at only 38 words, we cogitate that the parameter relating to the maximum of sequence length in the algorithms deployed in training can be potentially set at a lower commonly used bound of 256 rather than the maximum of 512 words to achieve the same accuracy. Whilst this is still significantly higher than the outlier threshold of 38, one has to take into account the additional length of the premise text and the potential of longer pieces of text during testing.

Furthermore, it is essential to examine the most important words in the subset and their differences to the original corpus. As can be observed in Figure 7, whilst the most essential words are still somewhat associated with climate as well as mitigation and adaptation practices, the themes present in the corpus seem to be more closely aligned to topics used in premises. We argue this is due to the fact that during the development of the dataset, more focus was put on including sentences that could inform the algorithm on how to differentiate small semantic nuances since a previous manual data exploration uncovered a large set of paragraphs that would talk about very similar themes from a

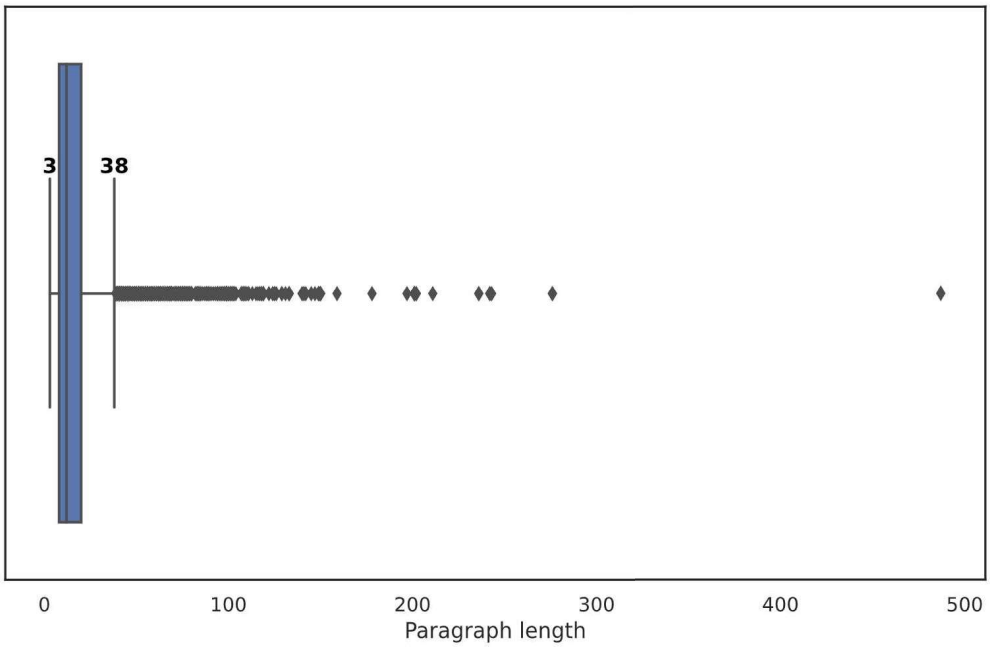


Figure 6: Boxplot of word count distributions for labelled dataset

Dataset	N	Y=1	Positive Fraction
Train	3193	1906	0.551
Evaluation	1064	619	0.565
Test	1064	590	0.538

Table 1: Count per label for each of the three datasets

semantic perspective, but would relate to different themes from the datapool of premises outlined earlier.

Lastly, we investigated the sample distribution of the true labels in order to understand what the appropriate evaluation metric is to deploy during training. As can be deduced from Table 1, the class labels are relatively evenly distributed across the training, evaluation and test datasets, which enables us to use the accuracy score as an appropriate metric for evaluation purposes.

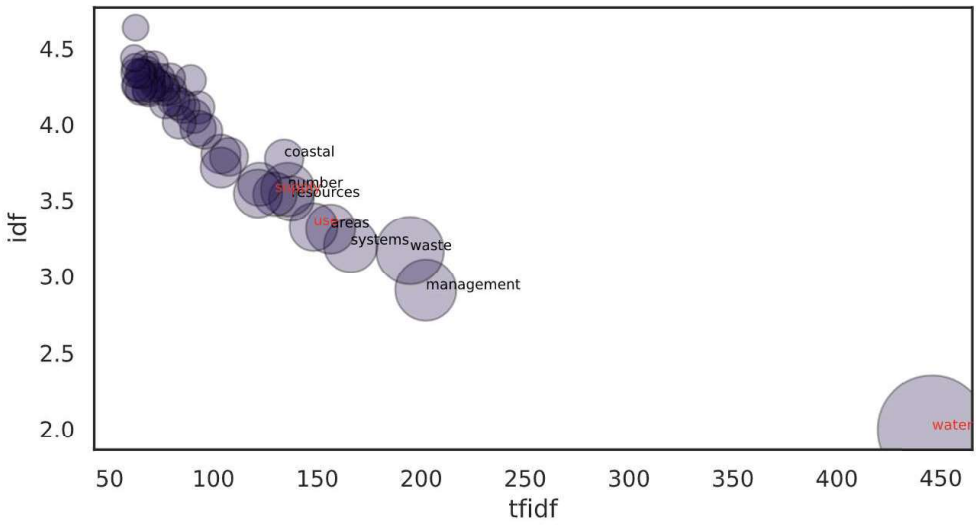


Figure 7: Most frequent words in labelled dataset

6 Model Definition

6.1 Algorithm

The P-BAAL algorithm can be derived by establishing a connection between its two core ideas, adapter based models and pool-based active learning. Following the same logic as when fine-tuning the adapter-based model, we start by incorporating adapter layers χ_v, according the into the pre-trained transformer Φ_w, such that the final model $\Psi_{w,v} = \chi_v(\Phi_w)$ and $|v| << |w|$. Subsequently, we "freeze" the layers of the base model w and initialise adapter weights v to near identity as performed by Houlsby et al., 2019.

Thereafter, following the approach of pool-based active learning, the first run of fine-tuning is initialised utilising only a subset of the train data pool T, where T can be defined as the sequence of observations X such that $\{X_1, \ldots, X_n\} = T \ \forall \ n > 0$. The initial subset used in training $train$ can in turn be specified as $\{X_1, \ldots, X_m\}_{x \in T} = train \ \forall \ 0 < m < n$. Additionally, the predictions of the remaining data in the pool $\{X_1, \ldots, X_k : k = n - m\}_{x \in T, x \notin train}$, or alternatively $U = T \cap train'$ will be stored for resampling in subsequent iterations. For each of prediction $\{f(x)_1, \ldots, f(x)_k\}$ returned by the model, we save their underlying probability classes and subsequently, calculate prediction entropies

$\{e_1, \ldots, e_k\}$ according to Equation 2 outlined in Section 3.4. This selection process follows the pool-based method of uncertainty sampling described in the same section. Finally, q instances characterised by the highest entropy will be incorporated into the training subset of data, $train$, before the fine-tuning is re-initialized with the original adapter weights. This approach is repeated until the model reaches the threshold of desired performance P, or the pool of test data U gets exhausted. Lastly, it is essential to note that the number of instances q subsequently added into the train set will follow a functional definition described later in this Section so that Hypothesis 3 can be investigated. The pseudocode for the P-BAAL algorithm is further described in Algorithm 1.

Algorithm 1 An algorithm with caption

Require:
$T \leftarrow \{X_1, \ldots, X_k\}$
$train \leftarrow \{X_1, \ldots, X_k\}_{x \in T}$
$U \leftarrow T \cap train'$
$0 < m < n$
$|v| \ll |m|$
$a > 0$
if $q = a$ **then**
$k_{max} \leftarrow 0$
else if $q = f(U_0)$ **then**
$k_{max} \leftarrow 256$
end if
$0 < P < 1$
Ensure:
$\Psi_{w,v} \leftarrow \chi_v(\phi_w(train)$
$\Psi_{w,v} \approx \phi_w(train)$
while $k > k_{max}$ and $accuracy < P$ **do**
Train $\Psi_{w,v}(train_t)$ for E epochs
Calculate $\{p_1, \ldots, p_k\} \leftarrow \{f(X_1), \ldots, f(X_k)\}$
Calculate $\{e_1, \ldots, e_k\} \leftarrow \{entropy(p_1), \ldots, entropy(p_1)\}$
$train_{t+1} \leftarrow train_t \bigcup \arg\max_q(\{e_1, \ldots, e_k\})$
Calculate $accuracy_t$ based on predictions $\{f(X_1), \ldots, f(X_k)\}$
end while

As discussed in Section 3.4, a potential alternative to the algorithm described above could be sampling using the query by committee (QBC) method. While we do not investigate in detail the performance differences between uncertainty sampling and QBC in

this research paper due to the high resource requirements of the latter, such a stacking approach could indeed further improve on the former approach through its inclusion of a potentially more meaningful informativity measure in form of the across model prediction variance.

6.2 Base model selection

Given that the predictive power of the final algorithm directly depends on the configuration of the pre-trained based model and its ability to appropriately capture the linguistic relationships in the dataset in use, we decided to conduct the fine-tuning experiments with a variety of pre-trained models provided by Hugging Face (Hugging Face, n.d.). It should be noted that all chosen models had to be supported by the Hugging Face' adapter-transformer extension so that the adapters could be utilised for fine-tuning purposes. This substantially limits the set of models that can be considered and prevents the utilisation of certain architectures, such as Mpnet (Song et al., 2020), which has shown to deliver superior performance for semantic search of paragraphs thanks to its ability to retain not only the dependency among predicted tokens, but also the full position information of a sentence.

While we opted to perform initial analyses with models characterised by rather simple designs, we subsequently extended the experiments with transformers carrying more complex architectures. Therefore, the models put as first under investigation include fairly simple pre-trained bidirectional transformers, namely bert-base-uncased (Devlin et al., 2019) and its distilled counterpart distilbert-base-uncased (Sanh et al., 2019). Subsequently, considering that the underlying language of the dataset in use is highly topic specific, relating to topics of climate change as well as climate mitigation and adaptation, we introduced architectures that have been additionally pre-trained on research papers and newspapers articles with climate-central themes in order to try to minimise the domain shifting problem present with transfer learning. In detail, this includes nbroad/ESG-BERT

(Pothireddi, C. and Parabole.ai, (2020), 2022) and climatebert/distilroberta-base-climate-f (Webersinke et al., 2021). Additionally, since specific model designs, such as sentence transformers based on architectures of MiniLM (Wang et al., 2020) , roberta-base (Liu et al., 2019) and its distilled counterpart distilroberta-base (Sanh et al., 2019), have previously shown to score well on tasks related to semantic search (Reimers and Gurevych, 2019), we extend our experimentation to encompass these model designs. The incorporation of all these models and their respective results will be shown in detail for the investigation of Hypothesis 1, whereas we will focus solely on the three models with the most interesting results in order to derive a conclusion for Hypotheses 2 and 3.

6.3 Adapter selection

To implement adapter layers into the base transformer model, we exploit the aforementioned Hugging Face' adapter-transformer extension (Pfeiffer et al., 2020). Whilst the library provides solutions designed to enable incorporation of multitude adapter structures, we limit our investigation to three designs. Firstly, we investigate performance of base models with simple bottle-neck adapters. The results are compared to experiments with more complex structures, namely with prefix-tuning, which has shown to perform well in low-resource scenarios. Lastly, we deploy mix-and-match (MAM) adapters, which are expected to deliver superior performance given their ability to effectively combine the essential features of parallel bottleneck adapters and prefix tuning. It should be noted that all investigations have been performed with the base adapter configuration provided for each adapter by Hugging Face.

Finally, following the approach of Poth et al., 2021, we experimented with pre-trained adapters during our preliminary studies. Nevertheless, since all trialled designs obtained from the Hugging Face database significantly underperformed the adapters whose weights were initiated as identity, we decided to exclude pre-trained adapters from the analysis. Similarly to the selected models described earlier, we opt to select only the best-

performing adapter with the most interesting results for further analysis in answering Hypotheses 2 and 3.

6.4 Sampling increment parameter q

As described earlier in this Section, the sample size increment q at each iteration should be formally defined using a functional form. For the analyses carried out in Hypotheses 1 and 2 this represents a constant increment of $q = a$.

Subsequently, we put under investigation the potential performance improvements of a declining value of q as a function of the size of the initial remaining sample pool U in Hypothesis 3. Whilst several different functional forms could be assumed such as a linear definition, we decided to define q using an exponentially declining function in the constant fraction r of U, as shown in Equation 4. In such a way, we ensure that the algorithm initially receives a high number of additional informative samples, ameliorating potential overfitting issues early on. At the same time, such a definition ensures a slower learning process at later iterations, where only a few observations are added each time, thus allowing the model to tweak itself optimally to the remaining informative observations. However, it should be noted that when deploying q defined as a non-constant, exponentially decreasing function of U_0, we terminate the algorithm when the number of instances in U is smaller or equal to 256 to prevent infinite loops of active learning arising from the functional form of q.

$$q = U_0(1-r)^i r \; \forall \; 0 <= r <= 1 \tag{4}$$

6.5 Hyperparamter definition

While the positive impact of P-BAAL should hold irrespective of the particular choice of hyperparameters tested, it is nonetheless essential to properly define the relevant set of

parameters and tune them to achieve an optimal model that is utilisable in a production setting. To achieve both these countering objectives, we decided to firstly investigate Hypothesis 1 using a generic set of hyperparameters, and subsequently focus our tuning efforts to the analyses carried out in Hypotheses 2 and 2.

Since training any type of neural network, even adapters, implies relatively high time constraints, we decided to focus on optimising a small range of vital parameters that relate to learning rate, input sequence length and number of epochs. Considering the learning rate α, we compared performances when a small rate of $\alpha = 0.00004$ is deployed as opposed to cases when a considerably higher rate of $\alpha = 0.0001$ is utilised. On the other hand, we decided to keep the batch size constant at a level 16.

Furthermore, a large portion of the experiments is dedicated to analysing the optimal parameter value for q to arrive at a conclusion for Hypothesis 2 and 3. We decided to set the initial constant increase $a = 256$. We believe this number to be a good initial hyperparameter, as it provides a significant increase in additional samples at each iteration, whilst at the same time ensuring that there are not too few active learning runs carried out due to q being close to the available sample size. In turn, we incorporate the set of options $a = \{128, 256, 512\}$ to provide a good understanding of the effect of the size of the parameter q in Hypothesis 2. Finally, the tunable parameter in question for Hypothesis 3 is the initial sampling pool fraction r which defines q through its exponential relationship in the sampling pool U_0. We believe the range of $r \in \{0.1, 0.15, 0.2, 0.3, 0.4\}$, to provide sufficient differences in potential sampling to properly understand its relationship on performance.

At the same time, it can be expected that the size of the initial training set $train$ has an effect on the resulting performance of the algorithm. Taking the above into considerations, exposing the model to too little data could lead to overfitting due to the model's convergence at a suboptimal local minimum and thus, the model would not be capable of rationally choosing new instances for training such that the knowledge gain is maximised. On the other hand, running an initial training on a sample size that is almost as large as

the total pool of labelled data would defeat the purpose of active learning, since the model would run out of options to additionally train on very quickly. Thus, taking the aforementioned concerns into account we trialled various initial sample size values in the range $\{0.1, 0.2, 0.3, 0.4\}$ of the original data pool size. The initial value for the training size during testing for Hypothesis 1 was set at 0.3.

Lastly, depending on the combination of adapter-transformer deployed and active learning parameter specification in use, we experiment with a number of training epochs to assure convergence. As the initial runs performed did not yet focus on finding the optimal set of model hyperparameters, but rather on identifying an existence of a possible pattern that could provide initial conclusions for Hypothesis 1, we let each model train for 7 epochs. Subsequent experimentations that focused on examining Hypotheses 2 and 3 explored a number of training epochs ranging from 7 to 10 to ensure an appropriate balance between performance optimization and overfitting prevention.

6.6 Evaluation method

As outlined in Section 5.4, we utilise a balanced training dataset for all experiments. Therefore, to assess and to compare performances across the trained models, we report the accuracy score, which is calculated on the test dataset. Additionally, for experiments that deploy active learning, we report on the number of training instances that were used when a model achieved its highest evaluation accuracy score. Finally, to understand whether a relationship exists between the rate at which additional data is added to the training dataset and the final predictive accuracy of the trained model, we report on Person correlation between the two variables.

7 Results

In this Section, we will provide a full overview of the results utilised to answer our research questions. We firstly concentrate on adapter-based models trained with a generic set of hyperparameters as defined in Section 6.5, for which we compare their performance when basic training is utilised to when P-BAAL is deployed. Subsequently a number of interesting base-model architectures for the highest performing adapter are investigated under the different hyperparameters values outlined. It should be noted that the results reported in parentheses throughout this Section will focus on the specific combination of model and parameters achieving the maximum results for each hypothesis under investigation only.

The results under the generic set of hyperparameters are summarised in Table 2. First of all, it can be deduced, that irrespective of the base model utilised, the MAM adapter (86.03) is significantly outperforming both the prefix-tuning (80.18), as well as bottleneck (76.71) adapters. This is an expected result given that MAM was elaborately designed to combine the strengths of prefix tuning and parallel bottleneck adapters. Additionally, we initially anticipated that due to the domain specific language present in the dataset in use, deploying base models additionally pre-trained on thematically related text corpora would ameliorate the predictive accuracy. Nevertheless, such phenomena has not been observed, considering that the best performance across all experiments is delivered by the bert-base-uncased base model (86.03), beating the domain specific models nboard/esg-bert (83.21) and climatebert (81.11) consistently, disrespective of the adapter design used or whether active learning was deployed. Therefore, it can be assumed that the issue of domain-shifting is not affecting the predictive power to a great extent and as a result, the general base BERT architecture has the ability to capture the textual semantics without the requirement for a full retraining of the model.

Furthermore, Table 2 provides interesting insights with respect to the inspection of the potential accuracy advancements that the incorporation of P-BAAL can bring. More pre-

	No Active Learning			Active Learning		
Adapter Type	**MAM**	**Prefix tuning**	**Bottleneck**	**MAM**	**Prefix tuning**	**Bottleneck**
Model						
bert-base-uncased	85.38	80.18	77.08	**86.03 (2829)**[1]	80.18 (3085)	76.71 (3085)
climatebert/distilroberta-base-climate-f	78.9	75.53	72.42	**81.11 (3085)**	74.89 (3085)	71.87 (3085)
nbroad/ESG-BERT	83.11	77.17	74.25	**83.21 (2317)**	75.89 (2829)	73.4 (3085)
sentence-transformers/all-MiniLM-L6-v2	77.72	71.32	70.22	**77.9 (3085)**	71.69 (2573)	69.06 (3085)
sentence-transformers/all-distilroberta-v1	81	74.88	72.5	**81.6 (3085)**	72.71 (3085)	72.5 (3085)
roberta-base	80.64	74.15	73.21	**81.83 (2317)**	71.23 (3085)	72.9 (2829)
distilbert-base-uncased	78.36	73.05	73.07	**78.88 (2829)**	73.24 (3085)	71.87 (3085)

1. Number of training samples used to achieve maximum score n = (X)

Table 2: Overview of results across different adapter under active learning and the no-active learning baseline

cisely, P-BAAL outperforms the adapter-only baseline only for the MAM adapter, amounting up to an average of 0.78 percentage points across all base models, while exploiting an average of just 83.58% of the full train data. Additionally, the highest performance improvement achieved by P-BAAL can be observed for the domain-specific adaptation of BERT, climateBERT, at 2.12 percentage points, whereas ESG-BERT and robert-base are reaching their maximum performance at only 68% of the total dataset used.

On the other hand, contradictory results are obtained for experiments with prefix tuning and bottleneck adapters, where active learning does not show to bring any improvements and in fact, decreases the models' predictive ability in most cases. In detail, P-BAAL diminishes the accuracy by -0.92 and -0.63 percentage points on average per prefix tuning and bottle neck adapter respectively, all while using almost the full set of data (89.05%, 91.24% on average). Whilst it is challenging to arrive at conclusive justification of the observed trend, we believe that performance advancements brought by the deployment of active learning is conditional to an already well-performing underlying model design.

Since the desired performance enhancements were observed only with MAM adapter-based models, we exclude prefix tuning and bottleneck adapters from the subsequent

Epochs	7	8	9
Model			
bert-base-uncased	86.03	86.28	86.09
nbroad/ESG-BERT	83.21	83.74	83.58
climatebert/distilroberta-base-climate-f	81.11	80.73	81.74

Table 3: Overview of results different epochs for each base model

analysis. Additionally, given the high time constraint characteristic of iterative procedures such as active learning, we limit the scope of experiments to utilisation of 3 based models, for which further hyperparameter tuning will be performed. In detail, we opt for bert-based-uncased and nbroad/esg-bert due to their superior posterior predictive accuracy. Additionally, we include climatebert, the model benefiting most significantly from the P-BAAL method.

As outlined in Section 6.5, we experimented with multiple values for the number of training epochs. The initial analysis showed that whilst training the model for 8 epochs can sometimes bring an additional improvement over the baseline approach with 7, utilising 9 epochs generally leads to overfitting, as shown in Table 3. Therefore, our subsequent analysis for understanding Hypothesis 2 and 3 will be reported for models fine-tuned using 8 epochs. Additionally, as outlined in Section 6.5, the initial train size is expected to impact the model's ability to pick new train instances in subsequent runs and the impact could vary based on the ratio that defines the number of labelled data that is to be added for training. Therefore, the consequent results are always reported across the defined set of initial train data sizes.

First of all, Table 4, summarise results obtained for the three models under different values for the constant increment parameter q and initial train size. It can be deduced that P-BAAL prefers smaller data accruals when iteratively learning, thereby showing support for the claim in Hypothesis 2. In fact, when computing the correlation between q and the accuracy score, as shown in Table 5, we observe a strong negative correlation,

	bert-base-uncased			nbroad/ESG-BERT			climatebert		
Increment size: q	**128**	**256**	**512**	**128**	**256**	**512**	**128**	**256**	**512**
Initial train size: *train*									
0.1	**86.21**	**86.21**	85.75	84.57	**84.75**	84.2	82.56	81.82	**82.92**
	(3161)[1]	(2905)	(2393)	(2393)	**(2393)**	(2905)	(3245)	(2905)	**(2893)**
0.2	**86.67**	86.39	85.58	84.38	**84.47**	83.2	**82.47**	81.09	80.73
	(2867)	(2739)	(2739)	(2867)	**(2739)**	(2739)	**(3251)**	(1971)	(2827)
0.3	**86.31**	86.28	85.56	**84.29**	83.74	83.2	**82.1**	82.01	79.27
	(3213)	(3085)	(2573)	**(2957)**	(2573)	(2573)	**(2317)**	(2829)	(2061)
0.4	86.39	**87.03**	86.21	83.65	**83.84**	83.19	**82.42**	82	80.82
	(2534)	**(2918)**	(2918)	(2918)	**(2406)**	(2406)	**(3302)**	(2662)	(2918)

1. Number of training samples used to achieve maximum score n = (X)

Table 4: Overview of results across different constant values for q and the initial size of $train$

Hypothesis	**H2**	**H3**
Model		
bert-base-uncased	-69.55 %	-49.62 %
nbroad/ESG-BERT	-56.67 %	-26.05 %
climatebert/distilroberta-base-climate-f	-47.96 %	-52.9 %

Table 5: Overview of the correlations between the tuneable parameter that defines q and the accuracy score

represented by $\rho = \{-0.696, -0.567, -0.48\}$ for bert-based-uncased, nbroad/esg-bert and climate bert models respectively. As shown by these values, the impact of a smaller q is most profound for BERT, where the maximum difference in accuracy across all initial training sizes between $q = 128\ (86.67)$ and $q = 512\ (85.54)$ is at 1.04 percentage points.

Nevertheless, it can be also observed that the models sometimes underperform when the smallest increment of $128\ (86.67)$ is utilised to cases when the increment amounting to $256(87.03)$ is used. This phenomena can potentially be attributed to the fact that the model might require to incorporate a larger proportion of new instances in early training iterations

	bert-base-uncased					nbroad/ESG-BERT					climatebert				
Increment as ratio of U: *r*	**0.1**	**0.15**	**0.2**	**0.3**	**0.4**	**0.1**	**0.15**	**0.2**	**0.3**	**0.4**	**0.1**	**0.15**	**0.2**	**0.3**	**0.4**
Initial train size: *train*															
0.1	**87.12**	86.3	86.57	86.57	85.21	**85.3**	84.75	84.29	85.21	83.93	**83.93**	83.47	83.28	81.73	79.74
	(2935)[1]	(2842)	(3188)	(2932)	(2784)	**(2249)**	(3134)	(2640)	(3199)	(2784)	**(3076)**	(3134)	(3121)	(2708)	(2905)
0.2	86.44	86.42	86.66	86.3	**86.72**	**85.38**	84.56	84.84	83.75	83.23	82.22	82.14	83.19	82	**83.1**
	(2673)	(2412)	(3084)	(3130)	**(3098)**	**(3116)**	(2910)	(3185)	(2991)	(2859)	(3150)	(2701)	(3158)	(2792)	**(3098)**
0.3	86.12	86.02	86.34	**86.68**	86.17	84.75	84.75	84.2	83.92	**84.93**	82.01	**83.11**	81.18	82.37	81.55
	(3050)	(2979)	(2663)	**(2626)**	(3142)	(2839)	(2895)	(3195)	(2626)	**(2933)**	(3090)	**(3111)**	(2948)	(3049)	(3142)
0.4	**87.21**	86.39	86.85	86.48	84.84	84.66	83.56	84.65	**85.2**	84.02	**82.28**	82.19	82.92	81.74	80.9
	(3026)	(3047)	(2776)	(2744)	(3008)	(2650)	(2790)	(3108)	**(3107)**	(2212)	**(3069)**	(3160)	(2606)	(3107)	(3008)

1. Number of training samples used to achieve maximum score n = (X)

Table 6: Overview of results across different values of the ratio r and the initial size of $train$

to prevent overfitting as this enables the model in later iterations to optimally choose the additional data to train on. At the same time, the table highlights that there does not seem to be any concrete between the accuracy and the initial training dataset size.

Second of all, in order to investigate the plausibility of Hypothesis 3, we focus on understanding whether defining the increment in the additional training sample as a declining function of the number of iterations could potentially increase the posterior predictive accuracy. Therefore, we perform active learning experiments utilising set of ratios r, defined according to logic outlined in Section 6.5, and compare the accuracies to the best performing combination of constant data increment and the initial train data size per base model trained with active learning, as reported in Table 6.

As can be deduced from Table 6, the incorporation of non-constant, exponentially decreasing data increment can further ameliorate the predictive accuracies, with the best enhancements amounting to 0.18, 0.55 and 0.99 percentage points per bert-based-uncased, nbroad/esg-bert and climatebert model respectively. While several of the parameter values for r across all initial train sizes actually seem to underperform with respect to the

baseline scores, with the maximum accuracy decrease equal to 3.18 percentage points, enough exploration of different values of r allowed us to identify a superior accuracy compared to any of the constant increment values for q set previously. This relationship holds for all models irrespective of the initial train dataset size. Moreover, the maximum result obtained for each of the three models, bert-base-uncased (87.21), nbroad/esg-BERT (85.38) and climatebert (83.93), during the entirety of our experiments can be found at $r = 0.1$. It seems that this value optimally sets the sample increment q as such that it initially learns from sufficient amounts of additional samples, while later focusing a lot of iterations on identifying small improvements. This result points towards a confirmation of research Hypothesis 3.

Overall, efficient tuning of the active learning parameters allows us to gain highly significant performance improvements on the mam-adapter only baseline. In fact, we achieved a total gain on test accuracy of 5.03 percentage points for climatebert, the model which already benefited the most from active learning under the baseline parameters. Furthermore, we achieved an accuracy gain of 2.27 and 1.83 percentage points for nbroad/esg-BERT and bert-base-uncased respectively. These results point towards support for the research hypothesis, specifically upon identifying the correct adapter and tuning for a non-constant increment of q.

8 Discussion

In this section, we provide a detailed discussion on the conclusions derivable from the results regarding the legitimacy of the proposed Hypotheses. This in turn enables us to answer the research question outlined in Section 1.

First of all, we initially hypothesised that the introduction of active learning into the fine-tuning stage of adapter-based models would enhance the performance over the adapter-only baseline. This claim can be partially supported when analysing the results obtained

in Section 7. In detail, P-BAAL brought significant increases in accuracy for all experiments with the MAM adapter, irrespective of the base model, initial train data size or the increments for data addition deployed. The opposite holds true however for experiments deploying prefix tuning or bottleneck adapters, where almost all experiments with P-BAAL underperformed compared to the baseline.

Since the MAM adapters substantially outperform the other two designs in the baseline experiments, we believe that the improvements gained via the introduction of P-BAAL are conditional to an already well-performing underlying model design. More precisely, base models with weak predictive capabilities might not be capable of identifying noisy data that when included for training, can bias model predictions much more significantly, given its high relative weight in the overall sample, as compared to when model is trained on millions of instances.

Additionally, the results point to the fact that the approach of P-BAAL seems to be more applicable to situations when the underlying transformer-based model has been additionally pretrained on some domain-specific corpus whose word distribution most likely differs from that of the dataset used in fine-tuning, e.g., the new dataset is more general or talks about slightly different topics. We believe that utilising P-BAAL in such scenarios enables the model to choose the right data instances characterised by semantic relationships, not otherwise encountered by the model during the pre-training phase and thus, helps to address the above distribution discrepancy. Considering all of the above, we conclude that the performance enhancement of P-BAAL is conditional to selecting the right combination of adapter module and base model that is well suited for learning on the data provided in the fine-tuning stage. As a result, we only partially accept Hypothesis 1.

Second of all, we tried to investigate the effect of the data increment q in P-BAAL on its predictive accuracy, hypothesising that allowing the model to include only the most informative instances for the subsequent iterations, can maximise the potential information gain. Indeed, when analysing the results, there seems to be a clear preference for

low values of data increments, which we further confirmed with the correlations obtained between the accuracy score and q for P-BAAL.

Nevertheless, some of the experiments showed to underperform in scenarios when the lowest value for the P-BAAL data increment was utilised. We believe that this phenomena arises due to the fact that if the model at early stage of P-BAAL is not only trained on small subset of data but also, is allowed to select a very sparse amount of additional instances for every subsequent run, the model will overfit to the initial subsample and thus, will lose the ability to strategically select the new data such that information gain is maximised. Moreover, the observed trend is more prevalent for experiments conducted with domain-specific base models, which are expected to prefer a higher rate for inclusion of new data instances in cases when the corpus used for additional pre-training might not fully coincide with the topics encountered in dataset utilised in fine-tuning.

In conclusion, while the above experiments did show some evidence for a preference of low values of q, some trials utilising the lowest q in combination with a domain specific base model underperformed. Therefore, we can assume that the model requires a larger amount of data in initial training iterations to lessen the effect of overfitting to the initially small subsample and thus, we can only partially accept Hypothesis 2.

In order to understand if the above assumption holds true, we dedicated the last part of the research to investigating whether defining q as an exponentially decreasing function of the fraction r of the initial sample pool U_0, rather than a constant number, could alleviate the overfitting issues. As outlined in Section 7, the approach has somewhat inferior results compared to the baseline for several experiments. Nevertheless, it was always possible to identify a value of r for which the model would beat the accuracy of all experiments trialled with a constant q, irrespective of the initial train dataset size deployed, thus providing support for Hypothesis 3, so long as the parameter is sufficiently tuned.

On top of that, the best results in this study were always obtained for the fraction $r = 0.1$, pointing towards the fact that a smaller initial fraction of U_0 is preferred. Indeed

such lower levels of r allow for the perfect combination of incorporating a sizable amount of additional data in early iterations and slowly learning on the few remaining observations later on.

These findings lead us to conclude that it is indeed possible to accept Hypothesis 3. Moreover, it can be argued that while P-BAAL already provides acceptable performance gains under the baseline parameter selection, the most significant gains can only be achieved when optimally tuning the parameter r such that the resulting sample increment q can be represented non-linearly. Hence, we believe the tuning stage of this parameter is essential for the successful application of P-BAAL.

All in all, the inference on the investigated hypotheses allows us to derive an answer regarding the research question outlined in Section 1. In detail, we conclude that introducing active learning to an adapter-based model can lead to performance enhancements on tasks characterised by resource constraints under two conditions. Firstly, active learning must be applied in combination with adapter architecture that delivers a strong performance even when deployed on its own. It should be noted that given the high time constraint characteristic of iterative training performed in active learning, it can be beneficial to perform basic fine-tuning on various adapter modules so that the most appropriate architecture for the dataset at hand can be identified beforehand. Secondly, the data increment q should be defined as a function of the initial sample pool U_0 that allows for inclusion of a higher proportion of new labelled instances in early runs, while restricting the model to choosing only few data points in the late stages of active learning.

If the two conditions are satisfied, P-BAAL delivers state-of-the-art performance when applied to tasks, which commonly suffer from resource constraints relating to low data availability and domain-specific language. As a result, P-BAAL has been successfully deployed in the production environment and is currently exploited by Climate Fund Managers to assign ESG scoring to incoming investment proposal pitches. Finally, this shows that P-BAAL helps to bridge the gap between the focus of academia, which concentrates on

developing theoretical NLP model concepts by performing rigorous testing on large corpuses of general text, and the need of business, which requires development of models capable of addressing the resource constraints arising from the deployment of custom datasets.

9 Conclusion and Future Work

In this research paper, we introduced P-BAAL: a pool based adapter active learning algorithm, which aims to address resource constraints relating to low data availability and domain specific language of the dataset in use. In order to assure that the theoretically developed model can be applied to a real business problem suffering from the above constraints, we evaluated the model's performance on a dataset currently exploited by an institutional investor for ESG assessment. Our experiments show that P-BAAL is capable of reducing the prediction error, when measured as an increase in accuracy as compared to baseline approach utilising adapters only, by up to 5.03 percentage points, amounting to an accuracy of 83.93 for the climatebert base model. At the same time, maximum accuracy was achieved for the base model BERT at 87.21, representing a 1.83 percentage point error reduction compared to the non-active baseline. Additionally, our approach delivers state-of-the-art performance and thus, we can confirm the research question described in Section 1. Hence, we believe that P-BAAL should be subject to further research to further ameliorate the potential performance capabilities.

In detail, we showed that the strong results are conditional upon selecting a suitable underlying adapter architecture that is capable of learning well on the dataset in use. In this study, we restricted our experiments to only a limited amount of adapter architectures due to the long execution time of iterative training of active learning. However, it could be interesting to see how the model performs under deployment of an alternative adapter architecture. Additionally, we exploit only base models, which are currently supported by

the adapter-transformer module of AdapterHub. This prevents the inclusion of models such as Mpnet, which have shown to perform especially well when deployed on natural language understanding tasks. Therefore, we believe further research should focus on investigating alternative combinations of adapters and base models to uncover even better performing designs.

On top of that, we showed that the performance can be further ameliorated when the model is allowed to choose a large proportion of labelled instances to train on in early iterations and is restricted to selection of only few instances in later stages. In this study, the rate q at which new labelled instances are included to the model during active learning iterations is defined as decreasing exponential function of size of the dataset available for training purposes. Nevertheless, we believe that given the high additional error reduction obtained from a non-constant definition of q, the further research should focus on experimentation on the functional definition of q such that would allow the model balance the high initial need for labelled data, while preventing it from inclusion of non-informative samples later on.

Lastly, we limit our research to utilisation of only one adapter-transformer architecture at once, which might not be capable of fully exploring the dataset space. To alleviate the issue, it would be interesting to understand whether a stacking of multiple models could further improve on the former approach through its inclusion of a potentially more meaningful informativity measure in form of the across model prediction variance. A similar idea motivates a query by committee (QBC) active learning approach, which trains a committee of models that will collectively decide on new data instances to be included in the pool for training. While we do not investigate in detail the performance differences between uncertainty sampling and QBC in this research paper due to the high resource requirements of the latter, we believe such a topic should be subject to further research.

References

Antoncic, M., Bekaert, G., Rothenberg, R. V., & Noguer, M. (2020). Sustainable investment-exploring the linkage between alpha, esg, and sdg's. *ESG, and SDG's (August 2020)*.

Bahja, M. (2021). Natural language processing applications in business. https://doi.org/10.5772/intechopen.92203

Bashar, M. A., & Nayak, R. (2021). Active learning for effectively fine-tuning transfer learning to downstream task. *ACM Trans. Intell. Syst. Technol.*, *12*(2). https://doi.org/10.1145/3446343

Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3615–3620. https://doi.org/10.18653/v1/D19-1371

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Ein-Dor, L., Halfon, A., Gera, A., Shnarch, E., Dankin, L., Choshen, L., Danilevsky, M., Aharonov, R., Katz, Y., & Slonim, N. (2020). Active Learning for BERT: An Empirical Study. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 7949–7962. https://doi.org/10.18653/v1/2020.emnlp-main.638

Faccia, A., Manni, F., & Capitanio, F. (2021). Mandatory esg reporting and xbrl taxonomies combination: Esg ratings and income statement, a sustainable value-added disclosure. *Sustainability*, *13*(16). https://doi.org/10.3390/su13168876

Grießhaber, D., Maucher, J., & Vu, N. T. (2020). Fine-tuning bert for low-resource natural language understanding via active learning. https://doi.org/10.48550/ARXIV.2012.02462

He, J., Zhou, C., Ma, X., Berg-Kirkpatrick, T., & Neubig, G. (2021). Towards a unified view of parameter-efficient transfer learning. https://doi.org/10.48550/ARXIV.2110.04366

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, *79*(8), 2554–2558. https://doi.org/10.1073/pnas.79.8.2554

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., de Laroussilhe, Q., Gesmundo, A., Attariyan, M., & Gelly, S. (2019). Parameter-efficient transfer learning for nlp. https://doi.org/10.48550/ARXIV.1902.00751

Hugging Face. (n.d.). Hugging Face Model Page.

Ikeagami, T., Kang, X., & Ren, F. (2022). Improvement of japanese text emotion analysis by active learning using transformers language model. *2022 14th International Conference on Computer Research and Development (ICCRD)*, 171–177. https://doi.org/10.1109/ICCRD54409.2022.9730387

Jacobs, B. I., & Levy, K. N. (2022). The challenge of disparities in esg ratings. *The Journal of Impact and ESG Investing*, *2*(3), 107–111. https://doi.org/10.3905/jesg.2022.1.040

Jiang, N., & de Marneffe, M.-C. (2019). Evaluating BERT for natural language inference: A case study on the CommitmentBank. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 6086–6091. https://doi.org/10.18653/v1/D19-1630

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, *28*, 11–21.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2019). BioBERT: A pre-trained biomedical language representation model for biomedical text mining (J. Wren, Ed.). *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btz682

Li, X. L., & Liang, P. (2021). Prefix-tuning: Optimizing continuous prompts for generation. https://doi.org/10.48550/ARXIV.2101.00190

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. https://doi.org/10.48550/ARXIV.1907.11692

Marsh, D. M. (1984). Michael hoey: On the surface of discourse. george allen and unwin, london. 211 pp. *Nordic Journal of Linguistics*, *7*(1), 77–79. https://doi.org/10.1017/S033258650000113X

Nugent, T., Stelea, N., & Leidner, J. L. (2020). Detecting esg topics using domain-specific language models and data augmentation approaches. https://doi.org/10.48550/ARXIV.2010.08319

OECD Development Assistance Committee. (2022). OECD DAC Rio Markers for Climate.

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. https://doi.org/10.3115/v1/D14-1162

Pfeiffer, J., Rücklé, A., Poth, C., Kamath, A., Vulić, I., Ruder, S., Cho, K., & Gurevych, I. (2020). AdapterHub: A framework for adapting transformers. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 46–54. https://doi.org/10.18653/v1/2020.emnlp-demos.7

Poth, C., Pfeiffer, J., Rücklé, A., & Gurevych, I. (2021). What to pre-train on? efficient intermediate task selection. https://doi.org/10.48550/ARXIV.2104.08247

Pothireddi, C. and Parabole.ai, (2020). (2022). Mukut03/ESG-Bert: Domain specific Bert model for text mining in sustainable investing.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018). Improving language understanding by generative pre-training.

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. https://doi.org/10.48550/ARXIV.1908.10084

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter. https://doi.org/10.48550/ARXIV.1910.01108

Schröder, C., Niekler, A., & Potthast, M. (2021). Revisiting uncertainty-based query strategies for active learning with transformers. https://doi.org/10.48550/ARXIV.2107.05687

Settles, B. (2010). Active learning literature survey.

Sokolov, A., Mostovoy, J., Ding, J., & Seco, L. (2021). Building machine learning systems for automated esg scoring. *The Journal of Impact and ESG Investing*, *1*(3), 39–50. https://doi.org/10.3905/jesg.2021.1.010

Song, K., Tan, X., Qin, T., Lu, J., & Liu, T.-Y. (2020). Mpnet: Masked and permuted pre-training for language understanding. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (pp. 16857–16867). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2020/file/c3a690be93aa602ee2dc0ccab5b7b67e-Paper.pdf

Taylor, W. L. (1953). “cloze procedure”: A new tool for measuring readability. *Journalism Quarterly*, *30*(4), 415–433. https://doi.org/10.1177/107769905303000401

United Nations for Climate Change. (n.d.-a). National Adaptation Programmes of Action.

United Nations for Climate Change. (n.d.-b). Nationally Determined Contributions (NDCs).

United Nations for Climate Change. (2020). National Adaption Plans 2020 in the formulation and implementation of NAPs.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems*. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Wang, W., Bao, H., Huang, S., Dong, L., & Wei, F. (2020). Minilmv2: Multi-head self-attention relation distillation for compressing pretrained transformers. https://doi.org/10.48550/ARXIV.2012.15828

Webersinke, N., Kraus, M., Bingler, J. A., & Leippold, M. (2021). Climatebert: A pretrained language model for climate-related text. https://doi.org/10.48550/ARXIV.2110.12010

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., ... Rush, A. (2020). Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6

A Appendix

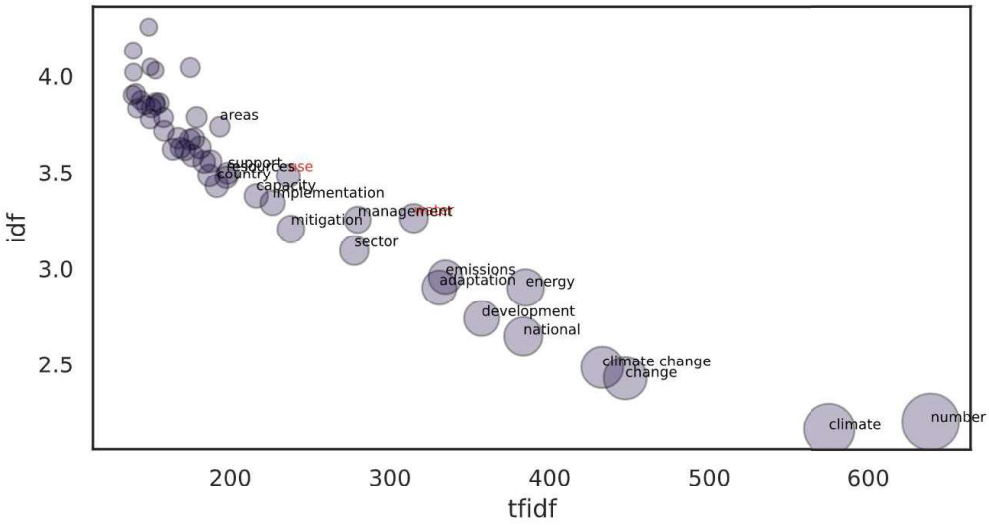


Figure 8: Most frequent words across NDC document

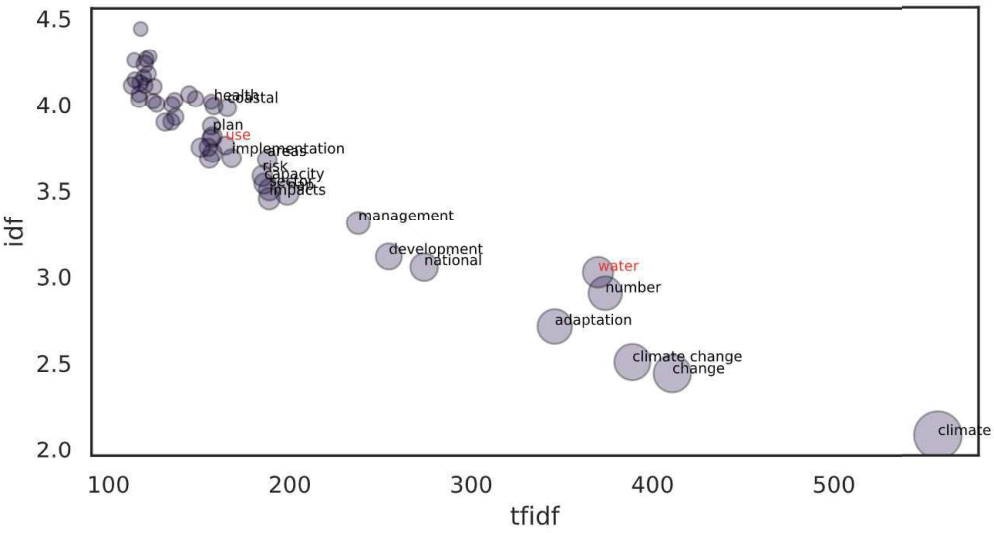


Figure 9: Most frequent words across NDC document

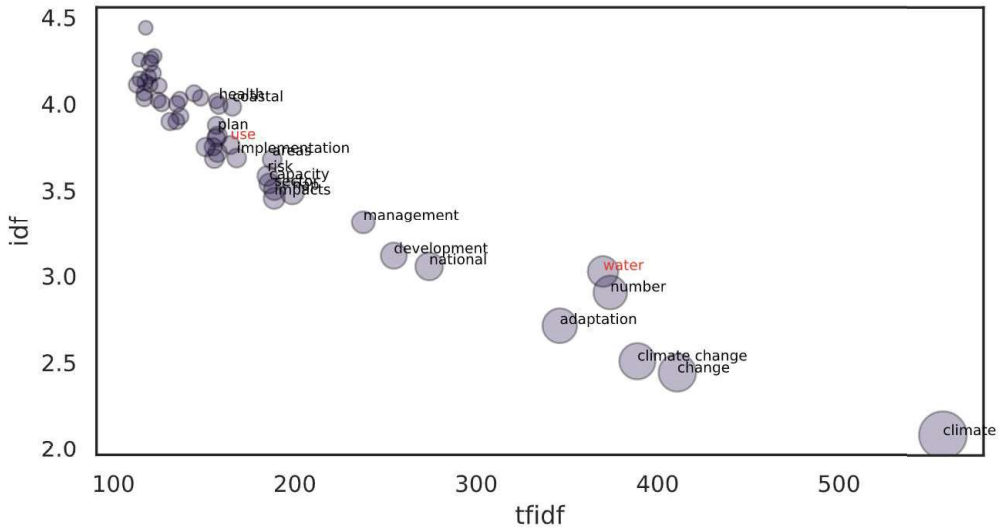


Figure 10: Most frequent words across NDC document