Active PETs: Active Data Annotation Prioritisation for Few-Shot Claim Verification with Pattern Exploiting Training

Xia Zeng, Arkaitz Zubiaga
Queen Mary University of London
{x.zeng,a.zubiaga}@qmul.ac.uk

Abstract

To mitigate the impact of the scarcity of labelled data on fact-checking systems, we focus on few-shot claim verification. Despite recent work on few-shot classification by proposing advanced language models, there is a dearth of research in data annotation prioritisation that improves the selection of the few shots to be labelled for optimal model performance. We propose Active PETs, a novel weighted approach that utilises an ensemble of Pattern Exploiting Training (PET) models based on various language models, to actively select unlabelled data as candidates for annotation. Using Active PETs for few-shot data selection shows consistent improvement over the baseline methods, on two technical fact-checking datasets and using six different pretrained language models. We show further improvement with Active PETs-o, which further integrates an oversampling strategy. Our approach enables effective selection of instances to be labelled where unlabelled data is abundant but resources for labelling are limited, leading to consistently improved few-shot claim verification performance.[1]

1 Introduction

As a means to mitigate online misinformation, research in automated fact-checking has experienced a recent surge of interest. Research efforts have resulted in survey papers covering different perspectives (Thorne and Vlachos, 2018; Kotonya and Toni, 2020; Nakov et al., 2021; Zeng et al., 2021; Guo et al., 2022) and novel datasets with enriched features (Augenstein et al., 2019; Chen et al., 2019; Ostrowski et al., 2021; Jiang et al., 2020; Schuster et al., 2021; Aly et al., 2021; Saakyan et al., 2021). Recent work has addressed various challenges, e.g. generating and utilising synthetic data (Atanasova et al., 2020; Pan et al., 2021; Hatua et al., 2021), joint verification over text and tables (Schlichtkrull et al., 2021; Kotonya et al., 2021), investigating domain adaptation (Liu et al., 2020; Mithun et al., 2021), achieving better evidence representations and selections (Ma et al., 2019; Samarinas et al., 2021; Si et al., 2021; Bekoulis et al., 2021), and performing subtasks jointly (Yin and Roth, 2018; Jiang et al., 2021; Zhang et al., 2021a).

As a core component of a fact-checking system, a claim validation pipeline consists of document retrieval, rationale selection and claim verification (Zeng et al., 2021). Our main focus here is claim verification, the task of assessing claim veracity with retrieved evidence. It is typically treated as a natural language inference (NLI) task: given a claim and an evidence, the aim is to predict the correct veracity label out of "Support", "Neutral" and "Contradict". Substantial improvements have been achieved in the performance of claim validation models when a considerable amount of training data is available (Pradeep et al., 2021; Li et al., 2021; Zeng and Zubiaga, 2021; Zhang et al., 2021b; Wadden et al., 2021). However, where new domains needing fact-checking emerge, collecting and annotating new relevant datasets can carry an impractical delay. Availability of unlabelled data can often be abundant, but given the cost and effort of labelling this data, one needs to be selective in labelling a small subset. In these circumstances, rather than randomly sampling this subset, we propose to optimise the selection of candidate instances to be labelled through active learning, such that it leads to overall improved few-shot performance.

To the best of our knowledge, our work represents the first such effort in proposing an approach leveraging an active learning strategy for the claim verification problem, as well as the first in furthering Pattern Exploiting Training (PET) with an active learning strategy. To achieve this, we propose Active PETs, a novel methodology that enables the ability to leverage an active learning strategy

[1]Our code is available here: https://github.com/XiaZeng0223/active_pets.

Findings of the Association for Computational Linguistics: EACL 2023, pages 190–204

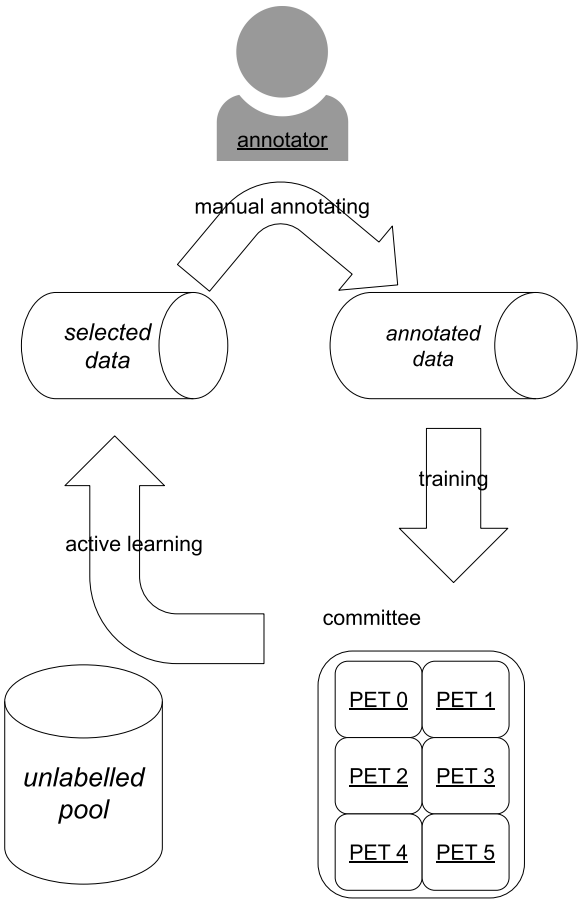

Figure 1: Illustration of the data annotation prioritisation scenario with a committee of 6 PETs. For each iteration, firstly the committee retrieves k new unlabelled samples (k=10 in our experiments), secondly the human annotators label them, lastly each of the PET based on different PLMs is trained individually with all of the labelled samples at hand. Our experiments start from 0 labelled samples and end at 300 labelled samples.

through a committee of PETs. Figure 1 illustrates the application of the active learning strategy on data annotation priotisation.

By exploring effective prioritisation of unlabelled data for annotation and making better use of a small amount of labelled data, we make the following novel contributions:

- we are the first to study data annotation prioritisation through active learning for few-shot claim verification;
- we are the first to study the extensibility of PET to enable active learning, by proposing Active PETs, a novel ensemble-based cold-start active learning strategy that enables multiple pretrained language models (PLMs) to collectively prioritise data instances;
- we further investigate the effect of oversampling on mitigating the impact of imbalanced data selection on few-shot learning, when guided by active learning;
- we conduct further corpus-based analysis on the selected few-shot data instances, which highlights the potential of Active PETs to lead to improved lexical and semantic characteristics that benefit the task.

Our results show consistently improved performance of Active PETs over the baseline active learning strategies on two datasets, SCIFACT (Wadden et al., 2020) and Climate FEVER (Diggelmann et al., 2021). In addition to improved performance over the baselines, our research emphasises the importance of the hitherto unexplored data prioritisation in claim verification, showing remarkable performance improvements where time and budget are limited.

2 Background

2.1 Claim Verification

Claim verification is typically addressed as an NLI problem (Thorne and Vlachos, 2018). Recent progress has enforced a closed-world reliance (Pratapa et al., 2020) and incorporated multiple instance learning (Sathe and Park, 2021). While data scarcity poses a major challenge on automated fact-checking (Zeng et al., 2021), research on few-shot claim verification is limited to date. Lee et al. (2021) investigated a perplexity-based approach that solely relies on perplexity scores from PLMs. Their model was tested on binary claim verification, as opposed to the three-way classification in our work. Zeng and Zubiaga (2022) introduced SEED, a vector-based method that aggregates pairwise semantic differences for claim-evidence pairs to address the task of few-shot claim verification. While their model addresses three-way classification, the experiments are only conducted in ideal scenarios where oracle evidences are available. To the best of our knowledge, however, no work has investigated the use of active learning in the context of claim verification. To further research in this direction, we propose Active PETs, a model that incorporates active learning capabilities into PET (Schick and Schütze, 2021a,b). PET has shown competitive performance in a range of NLP classification tasks, but its adaptation to the context of automated fact-checking and/or active learning settings has not been studied.

2.2 Active Learning

Active Learning (AL) is a paradigm used where labelled data is scarce (Ein-Dor et al., 2020). The key idea is that a strategic selection of training instances to be labelled can lead to improved performance with less training (Settles, 2009). Active learning methods are provided with an unlabelled pool of data, on which a querying step is used to select candidate instances to be annotated with the aim of optimising performance of a model trained on that data. The goal is therefore to optimise performance with as little annotation –and consequently budget– as possible. Traditional active learning query strategies mainly include uncertainty sampling, query-by-committee (QBC) strategy, error/variance reduction strategy and density weighted methods (Settles, 2012). Recent empirical studies have revisited the traditional strategies in the context of PLMs: Ein-Dor et al. (2020) examined various active learning strategies with BERT (Devlin et al., 2019), though limited to binary classification tasks. Schröder et al. (2022) conducted experiments with ELECTRA (Clark et al., 2020), BERT, and DistilRoBERTa (Sanh et al., 2019) respectively, while limiting the scope to uncertainty-based sampling.

Recent efforts on combining active learning with PLMs go into both warm-start and cold-start strategies. Warm-start strategies require a small initial set of labelled data to select additional instances, while cold-start strategies can be used without an initial set of labelled data. Ash et al. (2020) proposed Batch Active learning by Diverse Gradient Embeddings (BADGE) that samples a batch of instances based on diversity in gradient loss. Margatina et al. (2021) proposed Contrastive Active Learning (CAL), the state-of-the-art (SOTA) warm-start strategy that highlights data with similar feature space but maximally different predictions. Furthermore, Active Learning by Processing Surprisal (ALPS) (Yuan et al., 2020), the SOTA cold-start strategy, utilises masked language model (MLM) loss as an indicator of model uncertainty. We use BADGE, CAL and ALPS for baseline comparison, please see detailed descriptions in section 4.3.

To the best of our knowledge, QBC strategies (Seung et al., 1992; Dagan and Engelson, 1995; Freund and Haussler, 1997) that utilise a committee of models remains to be explored with PLMs, as previous studies limit their scope at measuring single model uncertainty. Nowadays various PLMs are publicly available that applying an ensemble-based query strategy on a downstream task becomes realistic, especially in few-shot settings where the computation required is relatively cheap. Furthermore, previous studies always perform fine-tuning to get classification results from PLMs. Our work presents the first attempt at integrating an active learning strategy into PET, which we investigate in the context of claim verification for fact-checking.

3 Methodology

In this section, we first describe PET, then introduce our model Active PETs, and finally describe the oversampling mechanism we use.

3.1 Pattern Exploiting Training

Pattern Exploiting Training (PET) (Schick and Schütze, 2021a,b) is a semi-supervised training procedure that can reformulate various classification tasks into cloze questions with natural language patterns and has demonstrated competitive performance in various few-shot classification tasks. To predict the label for a given instance x, it is first reformulated into manually designed patterns that have the placeholder $[mask]$. Then, the probability of each candidate token for replacing $[mask]$ is calculated by using a pretrained language model, where each candidate is mapped to a label according to a manually designed verbaliser.

3.2 Proposed method: Active PETs

Having a large pool of unlabelled data, our objective is to design a query strategy that selects suitable candidates to be labelled, such that the labelled pool of instances leads to optimal few-shot performance. Our query strategy is rooted in the intuition that disagreement among different PETs in a committee can capture the uncertainty of a particular instance.

Based on the assumption that performance of different language models is largely dependent on model size (Kaplan et al., 2020), we introduce a weighting mechanism: each PET is first assigned a number of votes V_i that is proportional to its hidden size,[2] and ultimately all votes are aggregated. Algorithm 1 presents the pseudo-code for executing a single query iteration with Active PETs.

[2]For example, if we use a committee formed of only base models that have 6 hidden layers and large models that have 12 hidden layers, proportionally each of the base models is allocated one vote and each of the large models is allocated two votes.

Algorithm 1 A Single Query Iteration

Require: The last trained Commitee of PETs C, unlabelled data pool U, query size k

for $PET_i \in C$ **do**
  $v_i \leftarrow Size(PET_i) / \min_{\forall PET_i \in C} Size(PET_i)$
end for ▷ assign number of votes
for instance $x \in U$ **do**
  for $PET_i \in C$ **do**
    $V_{x_i} \leftarrow resize(\hat{y}_{x_i}, v_i)$
  end for ▷ predict label and vote
  $S_x = -\sum_{\forall V_{x_i} \in V_x} \frac{V_{x_i}}{|V|} log \frac{(V_{x_i})}{|V|}$
end for ▷ calculate entropy scores
return $Sort(S)[:k]$ ▷ return top k instances

We then quantify the disagreement by calculating vote entropy (Dagan and Engelson, 1995):

$$score_x = -\sum_{\hat{y}} \frac{vote(x, \hat{y})}{count(V)} \log \frac{vote(x, \hat{y})}{count(V)} \quad (1)$$

where $\hat{y}$ is the predicted label, x is the instance, $vote(x, \hat{y})$ are the committee votes of $\hat{y}$ for the instance x, and $count(V)$ is the number of total assigned votes. It can be viewed as a QBC generalisation of entropy-based uncertainty sampling that is designed to combine models of different sizes.

3.3 Data Oversampling

One of the risks of the proposed active learning strategy is that the resulting training data may not be adequately balanced, which can impact model performance. An accessible solution is oversampling: resample the instances from the minority class with replacement until balanced. Note that this does not increase the labelling effort as instances are repeated from the labelled pool. Instead of random resampling (Japkowicz, 2000), we propose a novel technique of integrating resampling with the committee's preference. For each minority class, we start resampling from the instance that has the highest disagreement score to the instance that has the lower disagreement score. In highly imbalanced cases, resampling is repeated from the highest to lowest priority until the overall label distribution is balanced. Algorithm 2 presents the pseudo-code for executing the training loop with the option of conducting oversampling with Active PETs.

Algorithm 2 Training

Require: Labelled and sorted data D, A initial Commitee of PETs C

if Oversampling **then**
  $c \leftarrow max_{\forall class \in D} count(data \in class)$
  $D \leftarrow resize_{\forall class \in D}(class, c)$
end if ▷ oversampling
for $PET_i \in C$ **do**
  $PET_i \leftarrow train(PET_i, D)$
end for ▷ train the commitee of PETs
return C ▷ return trained PETs

4 Experimental Settings

Here we present the datasets and models used.

4.1 Datasets

SCIFACT			
	'Support'	**'Neutral'**	**'Contradict'**
UP	266 (9.31%)	2530 (88.55%)	61 (2.14%)
Test	150 (33.33%)	150 (33.33%)	150 (33.33%)
cFEVER			
	'Support'	**'Neutral'**	**'Contradict'**
UP	1789 (24.78%)	4778 (66.19%)	652 (8.66%)
Test	150 (33.33%)	150 (33.33%)	150 (33.33%)

Table 1: Label distribution of SCIFACT and cFEVER. UP = unlabelled pool of training data.

We choose real-world datasets with real claims, SCIFACT and Climate FEVER, known to be challenging, technical and free of synthetic data.[3]

SCIFACT provides scientific claims with their veracity labels, as well as a collection of scientific paper abstracts, some of which contain rationales to resolve the claims. In addition, it provides the oracle rationales that can be linked to each claim.

For SCIFACT, we perform the pipeline including abstract retrieval and claim verification. For the abstract retrieval step, we use BM25 to retrieve the top 3 abstracts, skipping the more specific rationale selection, as the SOTA system for this dataset suggested (Wadden et al., 2021). We chose BM25 based on high recall results reported in previous work (Pradeep et al., 2021). We merge original SCIFACT train set and dev set and redistribute the data to form a test set that contains 150 instances

[3] See data samples in Appendix A.

for each class and use the rest in the unlabelled pool. The reformulated data is highly imbalanced as presented in Table 1.

Climate FEVER (cFEVER) is a challenging large-scale dataset that consists of claim and evidence pairs on climate change, along with their veracity labels. As it does not naturally provide options of setting up retrieval modules, we directly use it on the task of claim verification. Similarly we reserve 150 instances for each class to form a test set and leave the rest in the unlabelled pool. Data in the unlabelled pool is heavily skewed, as shown in Table 1.

4.2 Active PETs

Committees of five to fifteen models are common for an ensemble-based active learning strategy (Settles, 2012). Here we form a committee of 6 PETs with 3 types of PLMs: BERT-base, BERT-large (Devlin et al., 2019), RoBERTa-base, RoBERTa-large (Liu et al., 2019), DeBERTa-base and DeBERTa-large (He et al., 2021). Given the commonalities between the NLI and claim verification tasks, we use the PLM checkpoints already fine-tuned on MNLI (Williams et al., 2018).

Despite a line of research in optimising PET patterns and verbalisers (Tam et al., 2021), that is not our main focus. We use the following pattern and verbaliser for PET: [claim]? *[mask]*, [evidence]; "Support":"Yes", "Contradict":"No", "Neutral":"Maybe", as they yielded best performance on NLI tasks in our preliminary experiments. Figure 2 provides an example of performing claim verification using PET.

There are two steps in our approach: (1) an ensemble method is used for data annotation prioritisation, after which data is selected and annotated, and (2) with the data instances drawn and annotated, we train a PET model that uses a single PLM to make the predictions. An ensemble method is key in step (1) to support the combined decision-making of choosing instances to annotate, but not in step (2) for the PET model which runs on a single PLM. Hence, results are presented for individual PETs, even if in all cases the ensemble is involved in the underlying prioritisation step. We test two variants: **Active_PETs** with no oversampling, and **Active_PETs-o** with the oversampling described in Section 3.3.

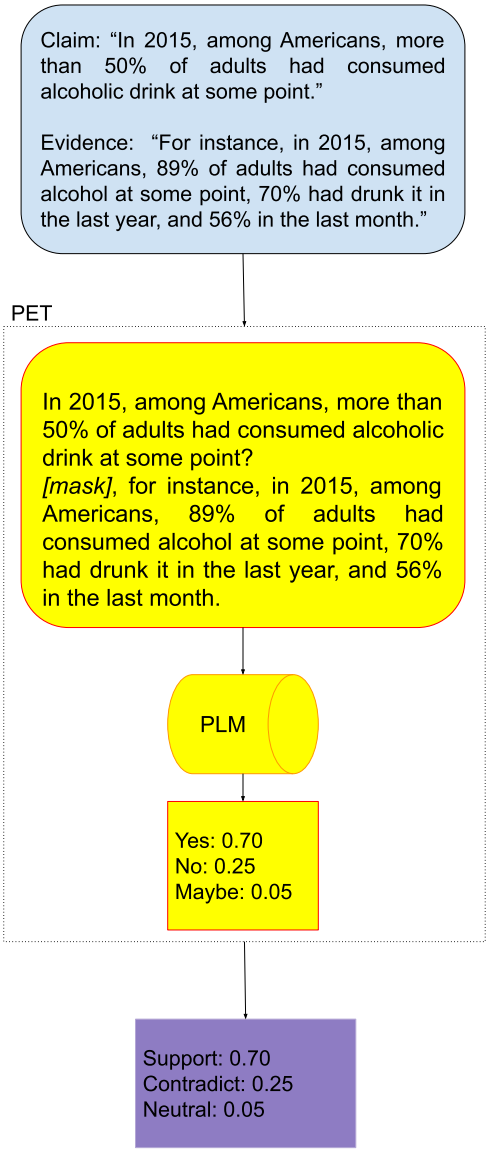

Figure 2: An example of doing claim verification with PET.

4.3 Baselines

We compare our method to four baselines: random sampling, BADGE, CAL and ALPS.

4.3.1 Random sampling

For random sampling, we run each experiment over 10 different sampling seeds ranging from 123 to 132, and present the averaged results.

4.3.2 BADGE

BADGE (Ash et al., 2020) optimises for both uncertainty and diversity. Gradient embeddings g_x are first computed for each data in the unlabelled pool, where g_x is the gradient of the cross entropy loss with respect to the parameters of the model's last layer. It then applies k-MEANS++ clustering

on the obtained gradient embeddings, and batch selects instances that differ in feature representation and predictive uncertainty.

Though BADGE is proposed as a warm-start method, the required initial set of labelled data is only used for the initial training the model. In our experiments on claim verification, PLMs that are already finetuned on a similar task NLI are used, hence, BADGE can be used for cold-start sampling.

4.3.3 CAL

CAL (Margatina et al., 2021), the SOTA warm-start strategy, highlights contrastive data points: data that has similar model encodings but different model predictions. Unlike BADGE, an initial labelled set of data is essential for CAL. It first calculates the [CLS] embeddings for all of the data and then runs K-Nearest-Neighbours (KNN) to obtain the k closest labelled neighbours for each unlabelled instance. It further calculates predictive probabilities from the model and measures Kullback-Leibler divergence on it. Finally it selects unlabelled instances whose predictive likelihoods diverge the most from their neighbours.

While CAL achieves SOTA performance as a warm-start strategy, its dependence on an initial labelled set of data makes it incompatible in the same few-shot active learning settings without an initial labelled set. However, for comprehensive comparison purposes, we still include it as a baseline starting at 100 labelled instances that are obtained from random sampling with 10 different random seeds.

4.3.4 ALPS

ALPS (Yuan et al., 2020), the SOTA cold-start active learning method, also aims to take both model uncertainty and data diversity into account. It calculates surprisal embeddings to represent model uncertainty. Specifically, for each instance x, it is passed through the masked language modelling head of a PLM and then 15% of the tokens in x are randomly selected to calculate the cross entropy against their target tokens. The surprisal embeddings go through L2-normalisation and then get clustered to select the top samples.

4.4 Training Details

Hyperparameters. As in few-shot settings we lack a development set, we follow previous work (Schick and Schütze, 2021a,b) and use the following hyperparameters for all experiments: $1e^{-5}$ as learning rate, 16 as batch size, 3 as the number of training epochs, 256 as the max sequence length. [4]
Labelling budget. We set it to a maximum of 300. We experiment with all scenarios ranging from 10 to 300 instances with a step size of 10.
Checkpoints. We always use the PLM checkpoints from the last iteration to perform active learning, but always train the initial PLMs which have never been trained on any fact-checking datasets.

5 Results

We next discuss results for our experiments.

5.1 Results on SCIFACT

Figure 3 presents experimental results on SCIFACT, where the unlabelled pool is large, heavily imbalanced and the domain is technical. Each subfigure shows results for a different PET among the six under consideration.

Data retrieved with Active PETs brings substantial improvements for all of the models, often from the very beginning but consistently as the number of shots increases from around 50 instances. Despite the performance fluctuations, training using data sampled with Active PETs rarely underperforms the baselines for SCIFACT. With Active PETs, Bert-base peaks at 0.352, RoBERTa-base peak at 0.345; DeBERTa-base peaks at 0.385; BERT-large peaks at 0.380; RoBERTa-large peaks at 0.409; DeBERTa-large peaks at 0.541. Generally, Active PETs shows a 10 to 20% increase in F1 scores, compared with various baselines.

Moreover, with Active PETs-o, i.e. when oversampling is further integrated with Active PETs, we observe a significant performance increase. Models tend to learn better from the beginning; the increase trend has less fluctuation; and the overall F1 scores are much higher. In this case, Bert-base peaks at 0.497, RoBERTa-base peak at 0.539; DeBERTa-base peaks at 0.551; BERT-large peaks at 0.548; RoBERTa-large peaks at 0.514; DeBERTa-large peaks at 0.587. This highlights the potential of oversampling, which increases the number of instances without additional labelling budget.

Among the baselines, we observe that training with data retrieved from all baselines failed to lead to any effective outcomes for BERT-base and DeBERTa-base within a labelling budget of 300 instances. While BADGE and CAL lead to some improvements over BERT-large and RoBERTa-large

[4] See further details for reproducibility in Appendix B.

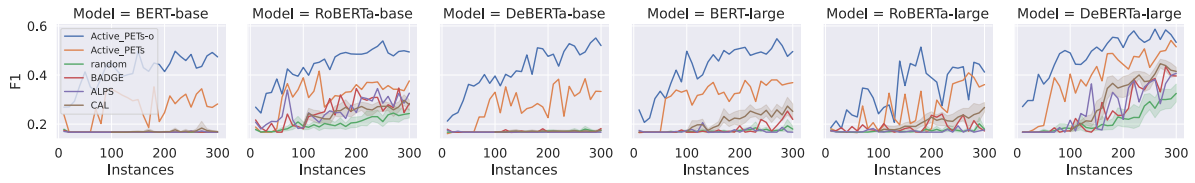

Figure 3: Few-Shot F1 Performance on SCIFACT claim verification.

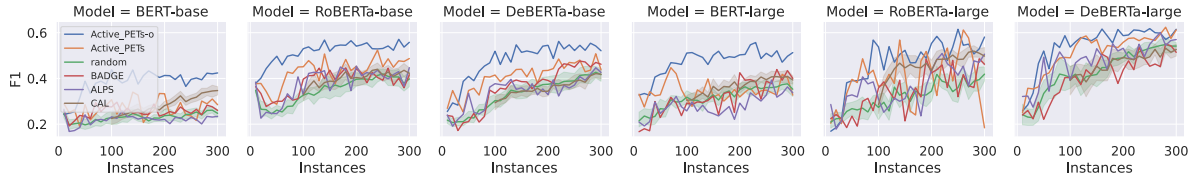

Figure 4: Few-Shot F1 Performance on cFEVER claim verification.

when given over 100 instances, random and ALPS failed to bring any improvements. Baseline results are better with RoBERTa-base and DeBERTa-large, but underperform Active PETs.

5.2 Results on cFEVER

Figure 4 presents F1 scores on cFEVER, where the unlabelled pool is large, imbalanced and the domain is somewhat technical. In this case, models generally achieve higher F1 scores than on SCIFACT. First of all, we observe that Active PETs outperforms random baseline in a more stable manner. It is over 10% higher than random most of the time, although it shows large performance fluctuations on RoBERTa-large. With Active PETs, Bert-base peaks at 0.34, RoBERTa-base peak at 0.524; DeBERTa-base peaks at 0.508; BERT-large peaks at 0.447; RoBERTa-large peaks at 0.612; DeBERTa-large peaks at 0.624. Moreover, Active PETs-o leads to a further performance boost, and more importantly, smooths out the large performance fluctuations. It is about 20% better than the random baseline most of the time. Specifically, Bert-base peaks at 0.438, RoBERTa-base peak at 0.571; DeBERTa-base peaks at 0.562; BERT-large peaks at 0.557; RoBERTa-large peaks at 0.615; DeBERTa-large peaks at 0.618.

When it comes to the baselines, the baselines do not struggle as much in the worst cases. Even if BERT-base's performance merely increased with most of the baselines, all of the other models managed to improve within the budget. With random sampling, RoBERTa-base, DeBERTa-base, BERT-large and RoBERTa-large all roughly peak at around 0.4, while DeBERTa-large is much better and peaks at around 0.5. BADGE, CAL and ALPS are in general better than random, but achieves lower F1 scores than Active PETs, especially in few-shot settings when the labelling budge is below 100.

6 Ablation Study

With SCIFACT we designed a slightly different pipeline where we conduct both evidence retrieval and claim verification – a setting that wasn't provided with cFEVER. To assess the impact of the addition of the evidence retrieval component on SCIFACT, we further perform ablation experiments on SCIFACT with oracle evidence.

With oracle evidence, the number of "Neutral" claim-evidence pairs are significantly reduced, resulting in a more balanced overall label distribution. After reserving 100 instances from each class for the test set, the unlabelled pool has 765 instances in total, where "Support" takes 46.54%, "Neutral" takes 38.43% and "Contradict" takes 15.03%. As shown in Figure 5, overall few-shot performance is much better and active learning demonstrates lesser performance gains. Sampling with baseline active learning strategies in general leads to similar results as random sampling. Surprisingly, coupling Active PETs with oversampling when the labelled pool is reasonably balanced, still maintains performance advantages across models. Under this setting, Bert-base peaks at 0.645, RoBERTa-base peak at 0.655; DeBERTa-base peaks at 0.766; BERT-large peaks at 0.68; RoBERTa-large peaks at 0.657; DeBERTa-large peaks at 0.86.

As demonstrated above, active learning is much

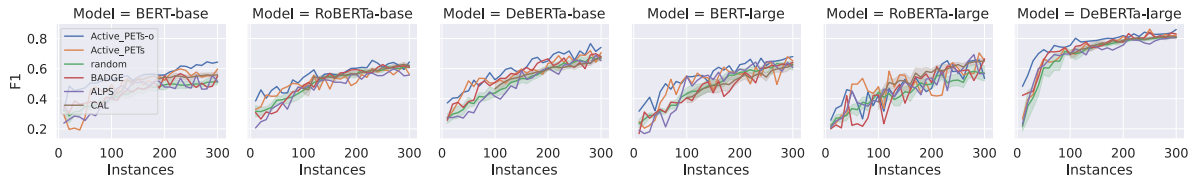

Figure 5: Few-Shot F1 Performance on Oracle SCIFACT claim verification.

more helpful for SCIFACT in a real-world setting than in an oracle setting. We could expect that if this finding generalises to cFEVER, active learning in a real-world setting involving evidence retrieval could possibly lead to larger performance gains.

7 Analysis

To better understand the impact of data prioritisation, we delve into the labelled data. In the interest of focus, we compare Active PETs with the SOTA cold-start method ALPS by analysing the best-performing PLM DeBERTa-large where 300 instances are selected.

7.1 Balancing Effects

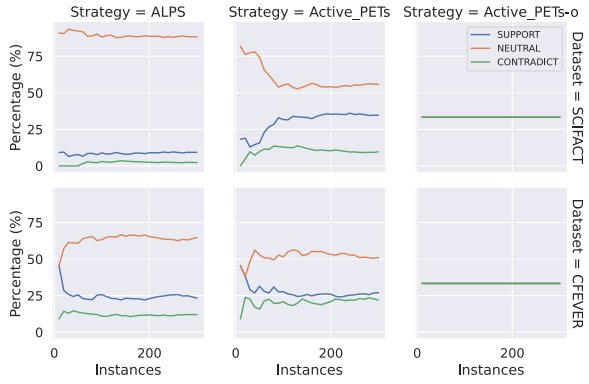

Figure 6: Label Distribution of data obtained with active learning by DeBERTa-large. The upper row is for SCIFACT and the lower row is for cFEVER.

We first look at the distribution of labels for the selected data. Figure 6 shows remarkable difference on label distribution for different active learning strategies. ALPS samples over 80% data from "Neutral", less than 10% from "Support" and very few from "Contradict" for SCIFACT; over 60% data from "Neutral", over 20% from "Support" and less than 20% from "Contradict" for cFEVER. They correlate well with original label distribution of each unlabelled pool, as presented in table 1. It suggests that ALPS is not sensitive to label distribution. However, Active PETs manages to sample a much more balanced distribution out of the extremely skewed original distribution. For SCIFACT, despite the initial few iterations, Active PETs samples less than 60% data from "Neutral", less than 40% data from "Support", around 10% data from "Contradict"; for cFEVER, Active PETs samples less than 60% data from "Neutral", over 20% data from "Support", around 20% data from "Contradict". In both datasets, label distribution from Active PETs are significantly more balanced than ALPS. Finally, the strategy with oversampling returns perfectly balanced distribution as expected. We identify a strong correlation between label distribution and classification performance.

7.2 Linguistic Effects

Aiming at providing further insights into data quality, we conduct corpus-based linguistic analysis to investigate lexical richness and semantic similarity.

Lexical Richness			
	ALPS	**Active_PETs**	**Active_PETs-o**
SCIFACT	0.0362	0.0387	0.0447
cFEVER	0.0389	0.0413	0.0503
Semantic Similarity			
	ALPS	**Active_PETs**	**Active_PETs-o**
SCIFACT	0.7921	0.8031	0.8054
cFEVER	0.7449	0.7744	0.7841

Table 2: Lexical richness is measured with Maas Type-Token Ratio (MTTR) scores and Semantic Similarity is measured by cosine similarity scores on embeddings of claims and evidences.

7.2.1 Lexical Richness

A popular metric for calculating lexical richness is Type-Token Ratio (TTR), where the total number of unique tokens is divided by the total number of tokens. We use Maas Type-Token Ratio (Maas TTR) (Maas, 1972), a logarithmic variant of TTR,

which is demonstrated to be less sensitive to the length of the text (McCarthy and Jarvis, 2007):

$$a^2 = \frac{logN - logV}{logN^2} \quad (2)$$

where N is the number of tokens in the corpus and V is the number of unique tokens in the corpus.

As shown in the upper part of Table 2, data selected by ALPS has the lowest lexical richness, while Active PETs leads to higher lexical richness for both datasets. Even more surprisingly, when integrating Active PETs with oversampling, the corpus has even higher score at lexical richness, despite that there are multiple duplicated instances in the corpus. One possibility is that training data with higher lexical richness may convey more useful information, as a bigger vocabulary enables more precise expressions.

7.2.2 Semantic Similarity

To investigate the overall data diversity, we calculate the average semantic similarity of all possible claim-evidence pairs in the corpus.[5] We obtain embeddings of claims and evidences with the PLM at interest, namely DeBERTa-large that has been trained on MNLI. For each embedded claim, we calculate its cosine similarity score with all embedded evidences in the corpus. The average of all similarity scores is then obtained. The lower part of Table 2 shows that ALPS leads to lowest overall semantic embedding similarity scores and Active PETs leads to higher scores. Integrated with oversampling, Active PETs leads to even higher similarity scores. It correlates well with the design of the strategies: ALPS explicitly encourages data diversity, while Active PETs focuses on committee uncertainty. One possible explanation is that data diversity is not as beneficial when the unlabelled pool contains less relevant instances: in the case of SCIFACT and cFEVER datasets, the majority of the unlabelled pool belongs to the "Neutral" class where the evidence is not enough to reach a verdict for the claim.

8 Conclusions

We present the first study on data annotation prioritisation for claim verification in automated fact-checking. With our novel method Active PETs, we demonstrate the potential of utilising a committee of PETs to collaboratively select unlabelled data for annotation, furthering in turn the extensibility of PET to active learning for the first time. Experiments on the SCIFACT and cFEVER datasets demonstrate the effectiveness of our proposed method, particularly in dealing with imbalanced data. Our proposed model consistently outperforms the random, BADGE, CAL and ALPS baselines by a margin. Further integration with an oversampling strategy that does not impact labelling effort leads to consistent performance improvements in all tested settings. Data that is more balanced shows to have higher lexical richness and semantic similarity, leading to better training results. While we have shown its effectiveness for claim verification here, in the future we aim to investigate Active PETs in other downstream tasks.

9 Limitations

We focus on demonstrating the effectiveness of Active PETs in scenarios where the labelling budget is limited and the label distribution is very imbalanced, as they are major challenges for automated fact-checking. Active PETs is shown to be particularly beneficial with low labelling budgets and becomes less so when the labelling budget increases and/or the unlabelled pool is balanced. Furthermore, as Active PETs is built on PET, it inherits the limitations from PET, e.g. a pattern-verbaliser pair (PVP) is required for any classification tasks. Note that a good selection of tested PVPs that cover common NLP tasks are publicly available.

Our experiments are only conducted with PLMs that are of base and large sizes, e.g., BERT-base and BERT-large, due to limited computing resources. Future work may further experiment with giant models like T5-11b and GPT-3. Another interesting direction would be to extend the proposed voting mechanism such that giant models and tiny models can both contribute effectively in the same committee, e.g., GPT-3 and DistillBert. Ideally, despite that GPT-3 is much larger than DistillBert, the extended voting mechanism should still allow DistillBert to contribute effectively.

Acknowledgements

We thank Christopher James Madge and Massimo Poesio from Queen Mary University of London for valuable pointers and comments; Ji-Ung Lee from Technische Universität Darmstadt for insightful discussions. Xia Zeng is funded by China Scholarship Council (CSC). This research utilised Queen

[5] Note that if we only calculate the retrieved pairs, the average similarity scores are approximately 1 for all strategies.

Mary's Apocrita HPC facility, supported by QMUL Research-IT. http://doi.org/10.5281/zenodo.438045

References

Rami Aly, Zhijiang Guo, Michael Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. FEVEROUS: Fact Extraction and VERification Over Unstructured and Structured information. *arXiv:2106.05707 [cs]*. ArXiv: 2106.05707.

J. T. Ash, Chicheng Zhang, A. Krishnamurthy, J. Langford, and Alekh Agarwal. 2020. Deep Batch Active Learning by Diverse, Uncertain Gradient Lower Bounds. *ICLR*.

Pepa Atanasova, Dustin Wright, and Isabelle Augenstein. 2020. Generating Label Cohesive and Well-Formed Adversarial Claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3168–3177, Online. Association for Computational Linguistics.

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. MultiFC: A Real-World Multi-Domain Dataset for Evidence-Based Fact Checking of Claims. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4685–4697, Hong Kong, China. Association for Computational Linguistics.

Giannis Bekoulis, Christina Papagiannopoulou, and Nikos Deligiannis. 2021. Understanding the Impact of Evidence-Aware Sentence Selection for Fact Checking. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 23–28, Online. Association for Computational Linguistics.

Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. 2019. Seeing Things from a Different Angle:Discovering Diverse Perspectives about Claims. In *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)*, pages 542–557, Minneapolis, Minnesota. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. *arXiv:2003.10555 [cs]*. ArXiv: 2003.10555.

Ido Dagan and Sean P. Engelson. 1995. Committee-Based Sampling For Training Probabilistic Classifiers. In *In Proceedings of the Twelfth International Conference on Machine Learning*, pages 150–157. Morgan Kaufmann.

J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2021. CLIMATE-FEVER: A Dataset for Verification of Real-World Climate Claims. *arXiv:2012.00614 [cs]*. ArXiv: 2012.00614.

Liat Ein-Dor, Alon Halfon, Ariel Gera, Eyal Shnarch, Lena Dankin, Leshem Choshen, Marina Danilevsky, Ranit Aharonov, Yoav Katz, and Noam Slonim. 2020. Active Learning for BERT: An Empirical Study. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7949–7962, Online. Association for Computational Linguistics.

Yoav Freund and David Haussler. 1997. Selective sampling using the query by committee algorithm. In *Machine Learning*, pages 133–168.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A Survey on Automated Fact-Checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Amartya Hatua, Arjun Mukherjee, and Rakesh Verma. 2021. Claim Verification Using a Multi-GAN Based Model. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 494–503, Held Online. INCOMA Ltd.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. *arXiv:2006.03654 [cs]*. ArXiv: 2006.03654.

Nathalie Japkowicz. 2000. The Class Imbalance Problem: Significance and Strategies. In *In Proceedings of the 2000 International Conference on Artificial Intelligence (ICAI*, pages 111–117.

Kelvin Jiang, Ronak Pradeep, and Jimmy Lin. 2021. Exploring Listwise Evidence Reasoning with T5 for Fact Verification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 402–410, Online. Association for Computational Linguistics.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. HoVer: A Dataset for Many-Hop Fact Extraction And Claim Verification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3441–3460, Online. Association for Computational Linguistics.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling Laws for Neural Language Models. *arXiv:2001.08361 [cs, stat]*. ArXiv: 2001.08361.

Neema Kotonya, Thomas Spooner, Daniele Magazzeni, and Francesca Toni. 2021. Graph Reasoning with Context-Aware Linearization for Interpretable Fact Extraction and Verification. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 21–30, Dominican Republic. Association for Computational Linguistics.

Neema Kotonya and Francesca Toni. 2020. Explainable Automated Fact-Checking: A Survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5430–5443, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Nayeon Lee, Yejin Bang, Andrea Madotto, and Pascale Fung. 2021. Towards Few-shot Fact-Checking via Perplexity. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1971–1981, Online. Association for Computational Linguistics.

Xiangci Li, Gully Burns, and Nanyun Peng. 2021. A Paragraph-level Multi-task Learning Model for Scientific Fact-Verification. *arXiv:2012.14500 [cs]*. ArXiv: 2012.14500.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692 [cs]*. ArXiv: 1907.11692.

Zhenghao Liu, Chenyan Xiong, Zhuyun Dai, Si Sun, Maosong Sun, and Zhiyuan Liu. 2020. Adapting Open Domain Fact Extraction and Verification to COVID-FACT through In-Domain Language Modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2395–2400, Online. Association for Computational Linguistics.

Jing Ma, Wei Gao, Shafiq Joty, and Kam-Fai Wong. 2019. Sentence-Level Evidence Embedding for Claim Verification with Hierarchical Attention Networks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2561–2571, Florence, Italy. Association for Computational Linguistics.

H.D. Maas. 1972. Uber den Zusammenhang zwischen Wortschatzumfang und Lange eines Textes. *Springer*, 8:73–96.

Katerina Margatina, Giorgos Vernikos, Loïc Barrault, and Nikolaos Aletras. 2021. Active Learning by Acquiring Contrastive Examples. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 650–663, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Philip M. McCarthy and Scott Jarvis. 2007. vocd: A theoretical and empirical evaluation. *Language Testing*, 24(4):459–488. Publisher: SAGE Publications Ltd.

Mitch Paul Mithun, Sandeep Suntwal, and Mihai Surdeanu. 2021. Data and Model Distillation as a Solution for Domain-transferable Fact Verification. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4546–4552, Online. Association for Computational Linguistics.

Preslav Nakov, D. Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, A. Barr'on-Cedeno, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated Fact-Checking for Assisting Human Fact-Checkers. In *IJCAI*.

W. Ostrowski, Arnav Arora, Pepa Atanasova, and Isabelle Augenstein. 2021. Multi-Hop Fact Checking of Political Claims. In *IJCAI*.

Liangming Pan, Wenhu Chen, Wenhan Xiong, Min-Yen Kan, and William Yang Wang. 2021. Zero-shot Fact Verification by Claim Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 476–483, Online. Association for Computational Linguistics.

Ronak Pradeep, Xueguang Ma, Rodrigo Nogueira, and Jimmy Lin. 2021. Scientific Claim Verification with VerT5erini. In *Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis*, pages 94–103, online. Association for Computational Linguistics.

Adithya Pratapa, Sai Muralidhar Jayanthi, and Kavya Nerella. 2020. Constrained Fact Verification for FEVER. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7826–7832, Online. Association for Computational Linguistics.

Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2021. COVID-Fact: Fact Extraction and Verification of Real-World Claims on COVID-19 Pandemic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2116–2129, Online. Association for Computational Linguistics.

Chris Samarinas, Wynne Hsu, and Mong Li Lee. 2021. Improving Evidence Retrieval for Automated Explainable Fact-Checking. In *Proceedings of the 2021 Conference of the North American Chapter of the*

Association for Computational Linguistics: Human Language Technologies: Demonstrations, pages 84–91, Online. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *ArXiv*.

Aalok Sathe and Joonsuk Park. 2021. Automatic Fact-Checking with Document-level Annotations using BERT and Multiple Instance Learning. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 101–107, Dominican Republic. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021a. Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021b. It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics.

Michael Sejr Schlichtkrull, Vladimir Karpukhin, Barlas Oguz, Mike Lewis, Wen-tau Yih, and Sebastian Riedel. 2021. Joint Verification and Reranking for Open Fact Checking Over Tables. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6787–6799, Online. Association for Computational Linguistics.

Christopher Schröder, Andreas Niekler, and Martin Potthast. 2022. Revisiting Uncertainty-based Query Strategies for Active Learning with Transformers. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2194–2203, Dublin, Ireland. Association for Computational Linguistics.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get Your Vitamin C! Robust Fact Verification with Contrastive Evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 624–643, Online. Association for Computational Linguistics.

Burr Settles. 2009. Active Learning Literature Survey. Technical Report, University of Wisconsin-Madison Department of Computer Sciences. Accepted: 2012-03-15T17:23:56Z.

Burr Settles. 2012. Active Learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 6(1):1–114.

H. S. Seung, M. Opper, and H. Sompolinsky. 1992. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, COLT '92, pages 287–294, New York, NY, USA. Association for Computing Machinery.

Jiasheng Si, Deyu Zhou, Tongzhe Li, Xingyu Shi, and Yulan He. 2021. Topic-Aware Evidence Reasoning and Stance-Aware Aggregation for Fact Verification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1612–1622, Online. Association for Computational Linguistics.

Derek Tam, Rakesh R. Menon, Mohit Bansal, Shashank Srivastava, and Colin Raffel. 2021. Improving and Simplifying Pattern Exploiting Training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4980–4991, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

James Thorne and Andreas Vlachos. 2018. Automated Fact Checking: Task Formulations, Methods and Future Directions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3346–3359, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or Fiction: Verifying Scientific Claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

David Wadden, Kyle Lo, Lucy Lu Wang, Arman Cohan, Iz Beltagy, and Hannaneh Hajishirzi. 2021. LongChecker: Improving scientific claim verification by modeling full-abstract context. *arXiv:2112.01640 [cs]*. ArXiv: 2112.01640.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Wenpeng Yin and Dan Roth. 2018. TwoWingOS: A Two-Wing Optimization Strategy for Evidential Claim Verification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 105–114, Brussels, Belgium. Association for Computational Linguistics.

Michelle Yuan, Hsuan-Tien Lin, and Jordan Boyd-Graber. 2020. Cold-start Active Learning through Self-supervised Language Modeling. In *Proceedings of the 2020 Conference on Empirical Methods*

in Natural Language Processing (EMNLP), pages 7935–7948, Online. Association for Computational Linguistics.

Xia Zeng, Amani S. Abumansour, and Arkaitz Zubiaga. 2021. Automated fact-checking: A survey. *Language and Linguistics Compass*, 15(10):e12438. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/lnc3.12438.

Xia Zeng and Arkaitz Zubiaga. 2021. QMUL-SDS at SCIVER: Step-by-Step Binary Classification for Scientific Claim Verification. pages 116–123.

Xia Zeng and Arkaitz Zubiaga. 2022. Aggregating Pairwise Semantic Differences for Few-Shot Claim Veracity Classification. ArXiv:2205.05646 [cs].

Zhiwei Zhang, Jiyi Li, Fumiyo Fukumoto, and Yanming Ye. 2021a. Abstract, Rationale, Stance: A Joint Model for Scientific Claim Verification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3580–3586, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhiwei Zhang, Jiyi Li, Fumiyo Fukumoto, and Yanming Ye. 2021b. Abstract, Rationale, Stance: A Joint Model for Scientific Claim Verification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3580–3586, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Example Appendix

We present example instances from SCIFACT and cFEVER datasets in this section.

B Reproducibility Appendix

We present further experimental details here for reproducibility.

Number of parameters in each model The number of parameters for BERT-base, BERT-large, RoBERTa-base, RoBERTa-large, DeBERTa-base, DeBERTa-large is 109484547, 335144963, 124647939, 355362819, 139194627 and 406215683 respectively.

Computing infrastructure We use High Performance Compute cluster supported by the university. Each experiment is run with 8 compute cores, 11G RAM per core and a single NVIDIA A100 GPU.

Run time Table 4 reports the average run time of executing a sampling iteration of 150 unlabelled instances and a training iteration with the sampled data over three datasets. It serves as a good indicator for comparing the efficiency among different active learning methods. As CAL requires an initial labelled set of data, we report the total run time of an iteration of using the random method for 75 instances and an iteration of using CAL method for another 75 instances. Table 5 further reports the total run time of the best method Active PETs-o on different datasets. The actual run time highly correlates with the size of the unlabelled pool for each datasets.

Our key focus has been on resource-efficiency and performance, with a lesser focus on runtime, hence there can be room for optimisation in future work, including: (1) optimising the code e.g. through parallelisation of the ensembled models which are now run sequentially, (2) using DL optimisation libraries such as deepspeed, and (3) using dynamic step sizes to reduce the number of iterations, e.g. increase step size if initial iterations lead to balanced samples. In a real-world, deployed scenario, one would also need to account for the time needed by humans to perform the annotation (in our case simulated).

SCIFACT		
Claim	**Evidence**	**Veracity**
"Neutrophil extracellular trap (NET) antigens may contain the targeted autoantigens PR3 and MPO."	"Netting neutrophils in autoimmune small-vessel vasculitis Small-vessel vasculitis (SVV) is a chronic autoinflammatory condition linked to antineutrophil cytoplasm autoantibodies (ANCAs). Here we show that chromatin fibers, so-called neutrophil extracellular traps (NETs), are released by ANCA-stimulated neutrophils and contain the targeted autoantigens proteinase-3 (PR3) and myeloperoxidase (MPO). Deposition of NETs in inflamed kidneys and circulating MPO-DNA complexes suggest that NET formation triggers vasculitis and promotes the autoimmune response against neutrophil components in individuals with SVV."	*"Suppport"*
"Cytochrome c is transferred from cytosol to the mitochondrial intermembrane space during apoptosis."	"At the gates of death. Apoptosis that proceeds via the mitochondrial pathway involves mitochondrial outer membrane permeabilization (MOMP), responsible for the release of cytochrome c and other proteins of the mitochondrial intermembrane space. This essential step is controlled and mediated by proteins of the Bcl-2 family. The proapoptotic proteins Bax and Bak are required for MOMP, while the antiapoptotic Bcl-2 proteins, including Bcl-2, Bcl-xL, Mcl-1, and others, prevent MOMP. Different proapoptotic BH3-only proteins act to interfere with the function of the antiapoptotic Bcl-2 members andor activate Bax and Bak. Here, we discuss an emerging view, proposed by Certo et al. in this issue of Cancer Cell, on how these interactions result in MOMP and apoptosis."	*"Contradict"*
"Incidence of heart failure increased by 10% in women since 1979."	"Clinical epidemiology of heart failure. The aim of this paper is to review the clinical epidemiology of heart failure. The last paper comprehensively addressing the epidemiology of heart failure in Heart appeared in 2000. Despite an increase in manuscripts describing epidemiological aspects of heart failure since the 1990s, additional information is still needed, as indicated by various editorials."	*"Neutral"*
Climate FEVER		
Claim	**Evidence**	**Veracity**
"In 2015, among Americans, more than 50% of adults had consumed alcoholic drink at some point."	"For instance, in 2015, among Americans, 89% of adults had consumed alcohol at some point, 70% had drunk it in the last year, and 56% in the last month."	*"Suppport"*
"Dissociative identity disorder is known only in the United States of America."	"DID is diagnosed more frequently in North America than in the rest of the world, and is diagnosed three to nine times more often in females than in males."	*"Contradict"*
"Freckles induce neuromodulation."	"Margarita Sharapova (born 15 April 1962) is a Russian novelist and short story writer whose tales often draw on her former experience as an animal trainer in a circus."	*"Neutral"*

Table 3: Veracity classification samples from the SCIFACT and Climate FEVER datasets.

	All Six Models	Average Single Model
Random	00:05:50	00:00:58
BADGE	00:07:52	00:01:19
CAL	00:14:59	00:02:30
ALPS	00:07:21	00:01:14
Active PETs	00:08:01	00:01:20
Active PETs-o	00:09:10	00:01:32

Table 4: Average run time for a single iteration for each of the sampling methods. The time format is hours:minutes:seconds.

	CFEVER	SCIFACT	Oracle SCIFACT
Active PETs-o	05:53:08	04:12:33	02:31:27

Table 5: Total run time for running active PETs with oversampling iteratively up to 300 instances on different datasets. The time format is hours:minutes:seconds.