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

Xia Zeng Arkaitz Zubiaga

School of Electronic Engineering and Computer Science
Queen Mary University of London

PANACEA workshop, September 2022

Table of Contents

1. Introduction
2. Background
3. Methodology
4. Results and Analysis
5. Conclusions
6. References

Table of Contents

1 Introduction

2 Background

3 Methodology

4 Results and Analysis

5 Conclusions

6 References

Claim verification: SCIFACT

	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."	*"Support"*
"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"*

Table: Veracity classification samples from the SCIFACT[1] dataset.

Claim verification: cFEVER

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."	*"Support"*
"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: Veracity classification samples from the Climate FEVER[2] dataset.

[2] Diggelmann et al. 2021.

Motivations

- Where new domains of misinformation emerge, collecting and annotating new relevant datasets can carry an impractical delay.
- Given the cost and effort of labelling this data, one needs to be selective in labelling a small subset.
- We propose to optimise the selection of candidate instances for overall improved few-shot performance.

Data annotation priotisation

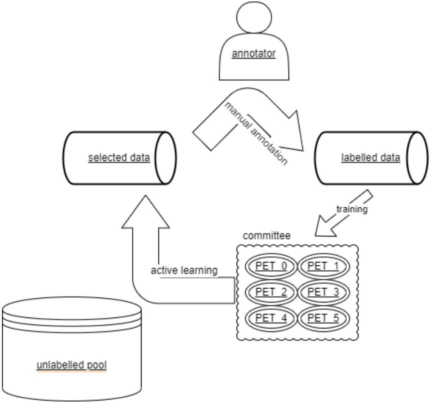


Figure: Illustration of the data annotation prioritisation scenario with a committee of 6 PETs.

Contributions

We present the following contributions:

- we are the first to study data annotation prioritisation for few-shot claim verification;
- we propose Active PETs, a novel ensemble-based cold-start active learning strategy;
- we further investigate the effect of oversampling on imbalanced data pool;
- we conduct further corpus-based analysis on the selected few-shot data instances.

Table of Contents

1. Introduction
2. Background
3. Methodology
4. Results and Analysis
5. Conclusions
6. References

Active learning query strategies

Classic strategies

uncertainty sampling, query-by-committee (QBC) strategy, error/variance reduction strategy and density weighted methods[a]

[a] Settles 2012.

Active learning query strategies

Classic strategies

uncertainty sampling, query-by-committee (QBC) strategy, error/variance reduction strategy and density weighted methods[a]

[a] Settles 2012.

Transformers era strategies

Batch Active learning by Diverse Gradient Embeddings (BADGE)[a], Active Learning by Processing Surprisal (ALPS)[b] *Contrastive Active Learning (CAL)*[c]

[a] Ash et al. 2020.
[b] Yuan, Lin, and Boyd-Graber 2020.
[c] Margatina et al. 2021.

Table of Contents

1. Introduction
2. Background
3. Methodology
4. Results and Analysis
5. Conclusions
6. References

Question:

How to utilise multiple PLMs for data annotation prioritisation, particularly in few-shot scenarios?

Pattern Exploit Training (PET)

PET achieves impressive few-shot performance on NLI;
it also makes many warm-start sampling strategies cold-start.

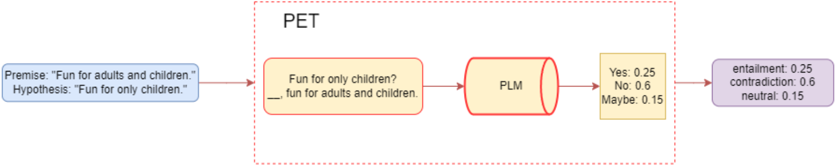

Premise: "Fun for adults and children."
Hypothesis: "Fun for only children."
PET
Fun for only children?
__, fun for adults and children.
PLM
Yes: 0.25
No: 0.6
Maybe: 0.15
entailment: 0.25
contradiction: 0.6
neutral: 0.15

Figure: An example of doing NLI with PET[3].

[3] Schick and Schütze 2021a; Schick and Schütze 2021b.

Proposed method: Active PETs

Assumption

Performance of different language models is largely dependent on model size[a].

[a] Kaplan et al. 2020.

Proposed method: Active PETs

Assumption

Performance of different language models is largely dependent on model size[a].

[a] Kaplan et al. 2020.

Weighting mechanism

Each PET is first assigned a number of votes V_i that is proportional to its hidden size and ultimately aggregate all votes.

Proposed method: Active PETs

Assumption

Performance of different language models is largely dependent on model size[a].

[a] Kaplan et al. 2020.

Weighting mechanism

Each PET is first assigned a number of votes V_i that is proportional to its hidden size and ultimately aggregate all votes.

Disagreement score with vote entropy:

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

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.

Proposed method: Active PETs-o

The unlabelled pool from both datasets are fairly imbalanced.

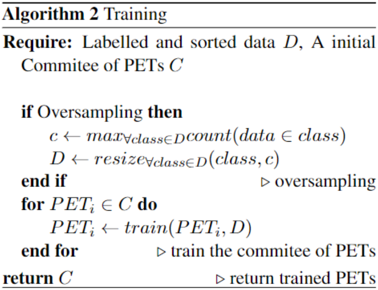
SCIFACT			
	‘Support’	**‘Neutral’**	**‘Contradict’**
UP	266 (9.31%)	2530 (88.55%)	61 (2.14%)
Test	150 (33.33%)	150 (33.33%)	150 (33.33%)

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: Label distribution of SCIFACT (with retrieved evidence) and cFEVER (with oracle evidence). UP = unlabelled pool of training data.

Proposed method: Active PETs-o

Even if we use active learning, imbalanced pool still leads to imbalanced samples. Hence, oversmapling:
'Support': 2, 'Contradict':3, 'Neutral':5 ⇒
'Support': 5, 'Contradict':5, 'Neutral':5



```
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
```

Figure: Training

Table of Contents

1 Introduction

2 Background

3 Methodology

4 Results and Analysis

5 Conclusions

6 References

SCIFACT

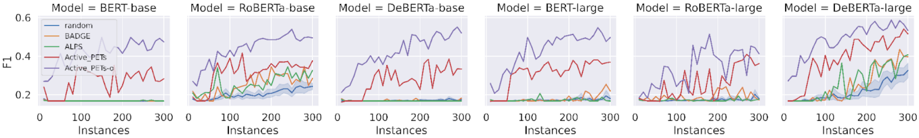


Figure: Few-Shot F1 Performance on SCIFACT claim verification. Evidences are the top 3 documents retrieved with BM25.

Active_PETs and Active_PETs-o achieve significant improvements.

cFEVER

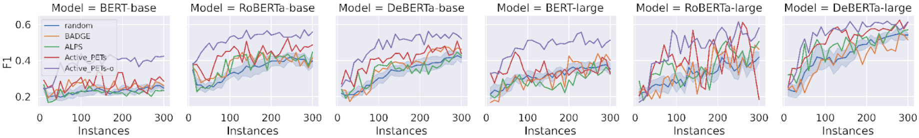


Figure: Few-Shot F1 Performance on cFEVER claim verification. Evidences are the oracle rationales as the cFEVER dataset does not offer document retrieval component.

Active_PETs and Active_PETs-o achieve significant improvements.

Balancing Effects

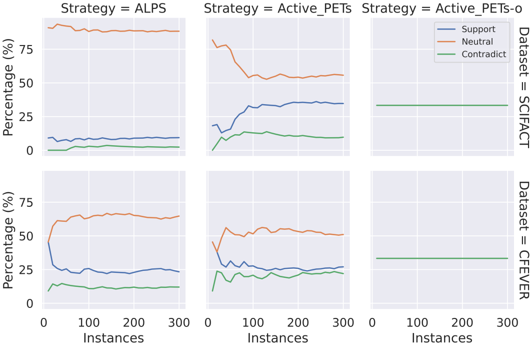


Figure: Label Distribution of data obtained with active learning with DeBERTa-large.

Linguistic Effects

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

Table of Contents

1 Introduction

2 Background

3 Methodology

4 Results and Analysis

5 Conclusions

6 References

Conclusions

- We present the first study in data annotation prioritisation for claim verification in automated fact-checking;

Conclusions

- We present the first study in data annotation prioritisation for claim verification in automated fact-checking;
- we demonstrate the effectiveness of Active PETs, particularly in dealing with imbalanced data;

Conclusions

- We present the first study in data annotation prioritisation for claim verification in automated fact-checking;
- we demonstrate the effectiveness of Active PETs, particularly in dealing with imbalanced data;
- Further integration with an oversampling component that doesn't impact labelling effort leads to consistent performance improvements;

Conclusions

- We present the first study in data annotation prioritisation for claim verification in automated fact-checking;
- we demonstrate the effectiveness of Active PETs, particularly in dealing with imbalanced data;
- Further integration with an oversampling component that doesn't impact labelling effort leads to consistent performance improvements;
- Data that leads to better training results is more balanced, has higher overall lexical richness, and higher semantic similarity within the pairs.

Table of Contents

1 Introduction

2 Background

3 Methodology

4 Results and Analysis

5 Conclusions

6 References

References I

Ash, J. T. et al. "Deep Batch Active Learning by Diverse, Uncertain Gradient Lower Bounds". In: *ICLR* (2020).

Diggelmann, Thomas et al. "CLIMATE-FEVER: A Dataset for Verification of Real-World Climate Claims". In: *arXiv:2012.00614 [cs]* (Jan. 2021). arXiv: 2012.00614. URL: http://arxiv.org/abs/2012.00614 (visited on 01/14/2022).

Kaplan, Jared et al. "Scaling Laws for Neural Language Models". In: *arXiv:2001.08361 [cs, stat]* (Jan. 2020). arXiv: 2001.08361. URL: http://arxiv.org/abs/2001.08361 (visited on 04/28/2022).

Margatina, Katerina et al. "Active Learning by Acquiring Contrastive Examples". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 650–663. DOI: 10.18653/v1/2021.emnlp-main.51. URL: https://aclanthology.org/2021.emnlp-main.51 (visited on 01/24/2022).

Schick, Timo and Hinrich Schütze. "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*. Online: Association for Computational Linguistics, Apr. 2021, pp. 255–269. DOI: 10.18653/v1/2021.eacl-main.20. URL: https://aclanthology.org/2021.eacl-main.20 (visited on 05/13/2022).

References II

Schick, Timo and Hinrich Schütze. "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*. Online: Association for Computational Linguistics, June 2021, pp. 2339–2352. DOI: 10.18653/v1/2021.naacl-main.185. URL: https://aclanthology.org/2021.naacl-main.185 (visited on 05/13/2022).

Settles, Burr. "Active Learning". en. In: *Synthesis Lectures on Artificial Intelligence and Machine Learning* 6.1 (June 2012), pp. 1–114. ISSN: 1939-4608, 1939-4616. DOI: 10.2200/S00429ED1V01Y201207AIM018. URL: http://www.morganclaypool.com/doi/abs/10.2200/S00429ED1V01Y201207AIM018 (visited on 01/13/2022).

Wadden, David et al. "Fact or Fiction: Verifying Scientific Claims". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 7534–7550. DOI: 10.18653/v1/2020.emnlp-main.609. URL: https://aclanthology.org/2020.emnlp-main.609 (visited on 05/13/2022).

Yuan, Michelle, Hsuan-Tien Lin, and Jordan Boyd-Graber. "Cold-start Active Learning through Self-supervised Language Modeling". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 7935–7948. DOI: 10.18653/v1/2020.emnlp-main.637. URL: https://aclanthology.org/2020.emnlp-main.637 (visited on 01/24/2022).