Reproducing Pattern-Exploiting Training (PET): A Critical Analysis of Few-Shot Text Classification Using AG News Dataset

First Author
Antara Tewary
G01413546
atewary@gmu.edu

Second Author
Ankit Kumar
G01436204
akumar37@gmu.edu

Third Author
Homa Haghighi
G01145449
hhaghigh@gmu.edu

1 Introduction

Schick and Schütze (2021) addresses the challenge of few-shot learning in NLP tasks, where only a small, labeled dataset is available. The research question is:
How can pre-trained language models (PLMs) like RoBERTa or BERT be adapted to excel at few-shot learning tasks using task-specific patterns and verbalizers?
The paper's main goal is to explore whether reformulating inputs as cloze-style questions can improve performance on tasks such as text classification and natural language inference (NLI) under low-resource conditions.
Link to our codebase (Tewary et al., 2024)

1.1 Task / Research Question Description

Core research questions are-
- Can natural language task descriptions be effectively combined with supervised learning to improve few-shot text classification?
- How can we exploit the implicit knowledge in pre-trained language models through carefully designed patterns?
- Is it possible to achieve better performance than standard supervised learning with as few as 10 labeled examples per class?

1.2 Motivation & Limitations of existing work

Existing few-shot learning methods often fail to effectively utilize pre-trained language models due to:

- *Lack of Task-Specific Adaptation*:Pre-trained models like BERT are not optimized for cloze-style tasks, which are effective for few-shot learning.
- *Heavy Dependence on Labeled Data*:Supervised models require large datasets, making them impractical for low-resource scenarios.

The proposed approach (PET) addresses these issues by:

- Exploiting task descriptions using patterns (templates) and verbalizers (label-to-word mappings).
- Enhancing semi-supervised training through iterative refinement (iPET), which improves performance by gradually increasing labeled data.

1.3 Proposed Approach

The core contribution is the Pattern-Exploiting Training (PET) framework:

- *Pattern and Verbalizer Design*:Reformulates tasks into cloze-style questions, enabling PLMs to better leverage their pre-training knowledge.
- *Soft Labeling*: Uses the model to annotate unlabeled data with probabilities (soft labels) to augment the training set.
- *iPET*: An iterative method that refines the labeled dataset over multiple training cycles, leading to improved generalization.

This approach enables few-shot learning by effectively combining human-readable patterns with the capabilities of large pre-trained models.

1.4 Likely challenges and mitigations

- *Computational Resources*
  Challenge: Training large models demands significant GPU power.
  Mitigation:Start with smaller models and subsets of patterns; use gradient accumulation.
- *Implementation Complexity*
  Challenge:Multi-stage training with intricate

interactions.
Mitigation: Modular design, extensive testing, and single-pattern trials.

2 Related Work

- Radford et al. (2019) introduced GPT-2, a language model capable of zero-shot learning by using task descriptions. However, it does not employ task-specific fine-tuning or semi-supervised learning, which PET utilizes to enhance few-shot performance.
- McCann et al. (2018) reframed NLP tasks as question-answering, enabling multitask learning. But it requires substantial labeled data, while PET leverages cloze-style patterns and verbalizers to use pre-trained models effectively in low-resource settings.
- Xie et al. (2020)presented a semi-supervised approach using data augmentation. PET differs by leveraging cloze-style patterns and pre-trained language models, avoiding reliance on data augmentation.
- Zeng and Zubiaga (2023) proposes "Active PETs", a weighted approach that uses an ensemble of PET models to actively select unlabeled data for annotation to improve few-shot claim verification. PET differs from traditional few-shot learning by reformulating tasks into cloze-style questions and using semi-supervised learning through iterative refinement to enhance performance with limited labeled data.

3 Experiments

3.1 Datasets

For this assignment, we focused on AGNews dataset, and its subset of four major news categories - World, Sports, Business, and Technology/Science. The original PET paper evaluated performance over multiple datasets like Yahoo Answers and Yelp Reviews, our focused approach enabled an in-depth analysis within our limited computational and time constrains.

3.2 Implementation

Our study (Tewary et al., 2024) builds upon the original implementation provided by the original paper, using their publicly available codebase as our foundation which ensures reproducibility and maintain direct comparability with the original work while conducting our experiments, architectural modifications, robustness analysis, and error studies.

- *Iteration 1* strictly follows the paper's configuration (100 training examples, batch sizes of 2/4, single epoch) to validate the reproducibility of the reported few-shot learning results. This is our baseline and allows us to verify PET's promise in low resource scenarios as claimed in the research.
- *Iteration 2* was done to explore beyond the paper's scope, deliberately increasing resources (200 training examples, batch sizes of 25/50, 3 epochs) to help us understand PET's behavior under enhanced conditions. This helps understand whether the orignial paper's conservative approach was optimal, or if PET's pattern-based approach could benefit from additional computational resources and training data.
- *Iteration 3* aimed to address some limitations within the paper.
  -Pattern Contextualization and Robustness:The original PET implementation used relatively simple patterns for news classification. Our enhanced pattern system introduces more natural, context-rich formulations like "What type of news is this?" and structured categorization patterns. This modification addresses a key limitation in the original paper-**the potential disconnect between pre-training objectives and downstream task formats**. With richer contextual patterns, we aim to better leverage RoBERTa's pre-trained knowledge of language structure and domain understanding.
  -AGNews-Specific Optimizations:A critical observation from the first two iterations was that the generic implementation might not fully exploit task-specific characteristics. The introduction of AGNewsConfig and specialized preprocessing demonstrates our hypothesis that news classification could benefit from domain-specific handling. The addition of special tokens like [DATA], [REPORT], and [NEWS] provides the model with explicit markers of document structure, potentially important for news article classification.
  -Data Quality and Representation:The enhanced preprocessing and metadata tracking

represent a significant shift from the original implementation's approach to data handling. By implementing thorough text cleaning and tracking features like headline length and numerical content, we address potential noise in news data that might affect model performance. This is particularly crucial in few-shot scenarios where each training example carries significant weight in model learning.

-*Verbalizer Validation and Flexibility*: The introduction of verifier mechanisms for single-token verbalizers and expanded label mappings (e.g., mapping "World" news to multiple relevant tokens like "Global" and "News") addresses a subtle but important limitation in the original implementation. This change provides more robust and flexible label representation, potentially allowing the model to capture different aspects of each news category.

-**Note:**These modifications maintain the same training configuration as Iteration 1 (reproducing the paper's setup), isolating the impact of architectural changes from training dynamics. This allows us to evaluate whether sophisticated pattern-verbalizer architectures and domain-specific optimizations can enhance performance even under conservative training conditions.

3.3 Results

Table 1 displaying result comparison with the original paper.

3.4 Discussion

The first iteration successfully reproduced the paper's results (88% ±0.0 vs 88.3% ±0.1) with identical training conditions, validating PET's reproducibility. Our architectural modifications in Iteration 3 improved training efficiency (loss reduction from 0.0598 to 0.0404) while maintaining accuracy. Iteration 2 achieved higher accuracy (90%) with more resources but increased training loss (0.0667), highlighting a performance-efficiency trade-off.

Notable differences include our consistent test accuracy (±0.0 vs paper's ±0.1), likely due to our single-pattern approach versus the paper's multiple patterns. While this simplified implementation, it potentially reduced result variability. The results demonstrate PET's reproducibility and potential for enhancement, though careful consideration of resource-performance trade-offs is crucial in few-shot learning scenarios.

3.5 Resources

Discuss the cost of your reproduction in terms of resources: computation, time, people, development effort, communication with the authors (if applicable).

3.6 Error Analysis

- *Training Loss vs Accuracy Trade-off*:Iteration 3's enhanced patterns achieved the lowest training loss (0.0404 vs 0.0598 in Iteration 1) without improving test accuracy (88%), suggesting improved training efficiency doesn't necessarily translate to better generalization.
- *Resource Scaling Efficiency*: Doubling training data and tripling epochs in Iteration 2 yielded only 2% accuracy improvement (90%), indicating diminishing returns in resource-performance trade-offs.
- *Pattern Effectiveness*: While modified patterns improved training accuracy (94% vs 92%), test performance remained unchanged, suggesting pattern enhancement alone doesn't guarantee better generalization.

These findings indicate PET's performance bottlenecks may lie in the fundamental pattern-based approach rather than training optimization. Future work should focus on pattern ablation studies and category-specific performance analysis to better understand these limitations.

4 Robustness Study

Our robustness evaluation framework assessed the model's resilience to input perturbations across different iterations of our implementation, with particular focus on our best performing model (Iteration 2, 90% accuracy). We designed two types of perturbation tests:

- *Character-level Perturbation*: Random character replacements with 0.1 probability, simulating common text errors like typos and misspellings.
- *Word-level Perturbation*: Random word removals with 0.1 probability, testing model resilience to incomplete information

Metric	Paper Results ($\|T\| = 100$ for PET)	Iteration 1 ($\|T\| = 100$)	Iteration 2 ($\|T\| = 200$)	Iteration 3 ($\|T\| = 100$)
Test Accuracy	88.3% ±0.1	88% ±0.0	90% ±0.0	88% ±0.0
Training Examples	100	100	200	100
Training Loss	Not reported	0.0598	0.0667	0.0404
Training Set Accuracy (Initial)	Not reported	24%	19.5%	24%
Training Set Accuracy (Final)	Not reported	92%	94.5%	94%
Configuration	RoBERTa-large Multiple patterns 1 epoch	RoBERTa-large Pattern ID 0 1 epoch	RoBERTa-large Pattern ID 0 3 epochs	RoBERTa-large Modified patterns 1 epoch

Table 1: Comparison of experimental results with the original paper.
Notes: $|T|$ represents the number of training examples. Paper results shown are for PET method (not iPET or supervised baselines). Iteration 1 and 3 directly compare with paper's $|T| = 100$ results. Iteration 2 uses more training examples (200) than reported configurations in the paper.

We systematically evaluated robustness across different training configurations (—T— = 10, 20, 50, 100) to understand how model stability correlates with training data size.

4.1 Results of Robustness Evaluation

- *Baseline Performance*
  - Consistent 88% accuracy across different few-shot configurations (10-100 examples)
  - Matches paper's reported performance (88.3% ±0.1) for $|T|$ = 100
  - Achieves 90% accuracy with increased training data ($|T|$ = 200)
- *Robustness Results*
  - Original Accuracy: 88%
  - Character Noise Impact: Drops to 4% accuracy (95.5% performance degradation) - Word Dropout Impact: Complete failure (0% accuracy)
- *Analysis: Successful Cases*
  - Base Model Stability: The model shows remarkable consistency in performance across different training set sizes, maintaining 88% accuracy from $|T|$ = 10 to $|T|$ = 100.
  - Training Efficiency: Lower training loss in Iteration 3 (0.0404) compared to Iteration 1 (0.0598) suggests improved learning efficiency with our modifications.
- *Analysis: Critical Vulnerabilities*
  - Severe Character Sensitivity: The dramatic drop to 4% accuracy under character noise indicates brittle pattern matching rather than robust semantic understanding.
  - Complete Word-level Failure: Zero accuracy under word dropout suggests the model lacks ability to handle incomplete information.

These results show that while our implementation improves the paper's accuracy, the model struggles with robustness to input perturbations. Strong performance on clean data (88-90%) versus poor resilience to perturbations (0-4%) highlights brittle dependencies in pattern-exploiting training for few-shot learning.

4.2 Discussion

Our robustness analysis revealed key insights and challenges for evaluating NLP models' robustness in few-shot learning scenarios:

- *Key Challenges*:
  - Pattern Dependency Trade-off: While PET's pattern-based approach enables effective few-shot learning, it creates vulnerability to input perturbations, suggesting a fundamental tension between pattern exploitation and input robustness.
  - Evaluation Limitations: The dramatic performance drop from 88% to 0-4% under perturbations indicates current evaluation metrics may be too binary, missing nuanced degradation patterns.
- *Recommendations for Future Research*:
  - Pattern Design: Implement redundant patterns capturing similar semantic information and design patterns accounting for common text variations.
  - Testing Framework: Develop more nuanced evaluation metrics and expand perturbation types beyond character noise and word dropout. Consider domain-specific variations.
- *Future Directions*:
  - Adaptive pattern selection based on input quality
  - Hybrid approaches combining pattern-based and traditional learning

- Benchmarks specifically for few-shot learning robustness

5 Workload Clarification

- *Code Implementation and Reproduction*: Each team member independently implemented the code, reproduced the paper's results, and verified consistency across implementations.
- *Report Writing*: Team members drafted separate sections, then collaboratively refined them into a cohesive, clear final report.

6 Conclusion

Our study reproduced the key findings of Pattern-Exploiting Training (PET), validating its effectiveness for few-shot text classification. We matched the original 88% accuracy on AGNews with 100 examples, and achieved 90% with 200 examples. However, our robustness analysis uncovered critical vulnerabilities - the model dropped from 88% on clean data to 4% and 0% under character and word perturbations. This brittleness highlights a tension between PET's pattern-exploiting and the need for robust language understanding. The key takeaways are:

- PET is a reproducible and effective few-shot technique, but architectural refinements did not improve generalization.
- PET's pattern-based nature makes it highly susceptible to input perturbations, underscoring the need for more robust few-shot learning approaches.

References

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze questions for few shot text classification and natural language inference.

Antara Tewary, Ankit Kumar, and Homa Haghighi. 2024. Pet reproducibility study. https://github.com/StringAna/PET_Reproducibility_Study.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. 2020. Unsupervised data augmentation for consistency training. In *Advances in Neural Information Processing Systems*, volume 33, pages 6256–6268.

Xia Zeng and Arkaitz Zubiaga. 2023. Active PETs: Active data annotation prioritisation for few-shot claim verification with pattern exploiting training. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 190–204, Dubrovnik, Croatia. Association for Computational Linguistics.