# Effective Few-Shot Classification with Transfer Learning

**Aakriti Gupta**
Yahoo! Research

**Kapil Thadani**
Yahoo! Research

**Neil O'Hare**
Yahoo! Research

{aakriti.gupta,thadani,nohare}@verizonmedia.com

## Abstract

Few-shot learning addresses the the problem of learning based on a small amount of training data. Although more well-studied in the domain of computer vision, recent work has adapted the Amazon Review Sentiment Classification (ARSC) text dataset for use in the few-shot setting. In this work, we use the ARSC dataset to study a simple application of transfer learning approaches to few-shot classification. We train a single binary classifier to learn all few-shot classes jointly by prefixing class identifiers to the input text. Given the text and class, the model then makes a binary prediction for that text/class pair. Our results show that this simple approach can outperform most published results on this dataset. Surprisingly, we also show that including domain information as part of the task definition only leads to a modest improvement in model accuracy, and zero-shot classification, without further fine-tuning on few-shot domains, performs equivalently to few-shot classification. These results suggest that the classes in the ARSC few-shot task, which are defined by the intersection of domain and rating, are actually very similar to each other, and that a more suitable dataset is needed for the study of few-shot text classification.

## 1 Introduction

To deal with emerging topics, events or linguistic phenomena, text classification systems often need to accommodate new classes. Recently, there has been increasing interest in addressing this problem through the lens of few-shot learning, which focuses on learning from a small handful of examples rather than a large annotated dataset. Few-shot classification research in computer vision and natural language processing studies the problem of generalizing from classes with plentiful training data to previously unseen classes for which labeled data may not be widely available.

At the same time, there has been a widespread growth of interest in transfer learning for natural language processing, with transformer-based self-supervised models such as BERT (Devlin et al., 2019) and T5 (Raffel et al., 2019) achieving strong results on a variety of benchmark problems. These models can be fine-tuned to new tasks with a small number of training examples, suggesting that their generalization capabilities may also be applicable to the few-shot learning setting.

In this paper, we study a straightforward transfer learning approach applied to a standard few-shot text classification dataset. Specifically, we focus on the well-studied Amazon Review Sentiment Classification (ARSC) dataset, which was originally developed for domain adaptation research (Blitzer et al., 2007), and was adapted to the few-shot setting by Yu et al. (2018). Various few-shot learning approaches have been benchmarked on this dataset, including techniques based on metric learning (Vinyals et al., 2016) and meta-learning (Finn et al., 2017). In this work, we consider a simple BERT-based classification scheme that first fine-tunes a pretrained model on the full rating classification dataset, and then further fine-tunes on only the held-out few-shot classes. This approach achieves comparable performance to state-of-the-art techniques, suggesting that pretrained models can extend their generalization capabilities to few-shot settings. Surprisingly, however, we also find similar performance in *zero-shot* settings for

*Proceedings of the 28th International Conference on Computational Linguistics*, pages 1061–1066
Barcelona, Spain (Online), December 8-13, 2020

this dataset, implying that few-shot categories are not sufficiently distinct from the other categories, and consequently motivating the need for new datasets to support future research in this area.

The contributions of our work are as follows:

- We examine the effectiveness of a simple BERT-based transfer learning strategy for few-shot classification, which makes use of class label prefixes to condition the model appropriately.
- We demonstrate the limits of the synthetic class distinctions in the ARSC few-shot dataset and discuss implications for future research on few-shot classification.

## 2 Related Work

The goal of few-shot learning (Miller et al., 2000; Fei-Fei et al., 2006; Wang et al., 2020) is to adapt a classifier to generalize to new classes using very few training examples. Such models typically cannot be trained using conventional methods, as modern classification algorithms require more parameters than there are training examples and this generalize poorly. Approaches based on *metric learning* attempt to relate new classes to those in the training data (Vinyals et al., 2016; Snell et al., 2017; Satorras and Bruna, 2018; Sung et al., 2018; Yu et al., 2018), while *meta-learning* techniques modify the optimization strategy to provide a model that can rapidly adapt to related tasks (Ravi and Larochelle, 2017; Finn et al., 2017; Mishra et al., 2018; Geng et al., 2019; Bansal et al., 2019; Deng et al., 2020). Recent work on both approaches has used the ARSC dataset (Blitzer et al., 2007), which we study in this work. We focus here on a straightforward application of the *transfer learning* paradigm popularized in natural language processing by models such as BERT (Devlin et al., 2019). These approaches exploit the language modeling capacity of transformer architectures (Vaswani et al., 2017) by pretraining models on large text corpora, and subsequently adapting these models to downstream tasks.

The link between few-shot learning and transfer learning is also suggested by prior work. Raffel et al. (2019) demonstrate that pretraining on in-domain unlabeled data can improve performance on downstream tasks, suggesting that the initial training data available to few-shot learners could also be used to improve generalization of pretrained models to few-shot classes. Brown et al. (2020) demonstrate that language models with multiple orders of magnitude more parameters than typically used in transfer learning can generalize to a wide variety of tasks with few training samples, implying that even domain mismatch is no barrier to the generalization ability of such models, given sufficient model capacity.

## 3 Dataset

In this work, we study the Amazon Review Sentiment Classification (ARSC) dataset used in recent few-shot learning literature (Yu et al., 2018; Geng et al., 2019; Deng et al., 2020). This dataset was originally constructed by Blitzer et al. (2007) as the Multi-Domain Sentiment Dataset[1] for the purpose of studying domain adaptation—a precursor task for few-shot learning—in sentiment analysis. The dataset consists of the text of reviews submitted to the Amazon online retail site for 25 product domains as well as their star ratings (out of 5), although 3-star reviews were excluded due to their ambiguous semantics for binary sentiment classification. Later work (Barzilai and Crammer, 2015) proposed alternative labels for these reviews based on the star rating $r$: $r = 5$, $r >= 4$ and $r >= 2$.

We use the version of the dataset adapted by Yu et al. (2018) for few-shot learning.[2] This version of the dataset employs the three rating categories proposed by Barzilai and Crammer (2015), drops stop-words and case from the review text, and partitions the corpus into training, validation and test splits. They propose a few-shot setting in which each domain and rating category is treated as a separate class. Four domains used in Blitzer et al. (2007)—BOOKS, DVDS, ELECTRONICS and KITCHEN & HOUSEWARES, which are among the largest domains in the dataset —are held out as few-shot domains, leading to a total of 12 few-shot classes (4 domains × 3 rating categories) for which training data is assumed to be scarce. Models are therefore able to use all training data from 57 classes (19 remaining domains[3] × 3 rating

[1] `https://www.cs.jhu.edu/~mdredze/datasets/sentiment/`

[2] `https://github.com/Gorov/DiverseFewShot_Amazon`

[3] There are 21 domains remaining but Yu et al. (2018) ignore the two smallest domains, which contain 444 examples in total.

categories) covering 95,586 training examples, but only 5 positive and 5 negative examples from each of the 12 few-shot classes. Trained few-shot classifiers are evaluated only on test sets for these 12 classes.

## 4 Model Formulation

We consider a straightforward application of transfer learning to the few-shot classification scenario described above, based on the original BERT model of Devlin et al. (2019), which employs a pretrained transformer encoder supervised by two objectives: masked token prediction and next sentence prediction.

Although domain labels in the ARSC dataset are expected to be disjoint, the rating categories are not, e.g., a review that has a positive label for $r = 5$ would also have a positive label for $r >= 4$ and $r >= 2$. Thus, more than one class label may be positive for a given input example, and so the learning problem is a form of multi-label classification—where examples can be assigned multiple class labels—rather than mutually-exclusive multi-class classification. We address this by formulating the prediction task as a binary classification problem, where the rating category and domain are encoded in the input along with the text to be classified, and the target label is a single binary label corresponding to that rating-domain combination. This is done by adding plain-text prefixes to each review, in training and during inference, to identify the star rating that the model is expected to distinguish and the domain that the data belongs to (e.g., "task two books" for $r >= 2$ and the BOOKS domain). So, instead of training a separate model for each rating-domain pair, we add this class information to the model input, and train a single binary model to learn across all classes. Prior work with transformer models has shown that such an approach can be highly effective, and prefix input tokens have been used to inject task information into models by specifying the desired output language or format (Raffel et al., 2019), domain (Keskar et al., 2019), etc.

Pretrained BERT models are typically fine-tuned on the target task directly; however, the extremely small number of training examples for few-shot classes (120 in total: 5 each for positive/negative examples across 12 rating-domain pairs) renders this approach unviable, as is also shown by Bansal et al. (2019). Therefore, in order to adapt the pretrained model to the few-shot learning domain, we conduct two rounds of fine-tuning: first with the 57 classes for which full training data is available, and then a second round of fine-tuning with only the 12 few-shot classes.

## 5 Experiments

We investigate the usefulness of this simple transfer learning approach through evaluations on the ARSC dataset. The BERT-base pretrained model released in the Transformers library[4] is used as a starting point and fine-tuned on the review data with sequence lengths truncated to 256 subword tokens. Two kinds of class prefixes are evaluated: the rating category alone (BERT-FS$_r$) and the rating-domain pair (BERT-FS$_{r+d}$). Comparisons are based on the mean per-class accuracy over test data following Yu et al. (2018) and subsequent work.

### 5.1 Hyperparameters

The first round of fine-tuning is conducted to convert the pretrained BERT model into an effective rating category predictor on the 57 classes in the dataset for which all training data is available. We tuned the learning rate and number of epochs on the validation set for this data and the best models were obtained within 10 epochs using a batch size of 16 and learning rates of 2e-6 (BERT-FS$_r$) and 5e-6 (BERT-FS$_{r+d}$).

Obtaining hyperparameters for the second round of fine-tuning is not straightforward, however, as the few-shot learning setting restricts the use of a validation set for the 12 few-shot classes. We follow the approach of Yu et al. (2018) and construct a surrogate few-shot learning dataset by selecting 12 of the 57 training classes (4 of the 19 training domains[5]) and keeping only 5 random positive and negative examples from each for fine-tuning. The validation data for these 12 classes is used to optimize hyperparameters and the remaining 45 training classes are used for first-round fine-tuning as above. In this hyperparameter search, we observe that fine-tuning with few-shot data leads to improvements of up to

[4] `https://github.com/huggingface/transformers`

[5] We select the 4 smallest domains in the training data for the surrogate few-shot dataset—AUTOMOTIVE, CELL PHONES SERVICE, GOURMET FOOD and OFFICE PRODUCTS—to ensure the first fine-tuning phase has sufficient training examples.

| Model | Accuracy (%) |
|---|---|
| Matching network (Vinyals et al., 2016) | $65.73^*$ |
| Prototypical network (Snell et al., 2017) | $68.15^*$ |
| MAML (Finn et al., 2017) | $78.33^\ddagger$ |
| RobustTC-FSL (Yu et al., 2018) | $83.12^*$ |
| SNAIL (Mishra et al., 2018) | $82.57^\dagger$ |
| Graph network (Satorras and Bruna, 2018) | $82.61^\dagger$ |
| Relation network (Sung et al., 2018) | $83.07^\dagger$ |
| Induction network (Geng et al., 2019) | $85.63^\dagger$ |
| MTM (Deng et al., 2020) | $\mathbf{90.01}^\ddagger$ |
| BERT-FS$_{r+d}$ | 89.59 |
| BERT-FS$_r$ | 88.99 |

Table 1: Mean accuracy on test dataset for 12 few-shot categories. Results from prior work reported by $^*$Yu et al. (2018), $^\dagger$Geng et al. (2019) and $^\ddagger$Deng et al. (2020).

| Model | BOOKS | | | DVDS | | | ELECTRONICS | | | KITCHEN & HOUSE | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $= 5$ | $>= 4$ | $>= 2$ | $= 5$ | $>= 4$ | $>= 2$ | $= 5$ | $>= 4$ | $>= 2$ | $= 5$ | $>= 4$ | $>= 2$ |
| 5-shot | 75.50 | 94.70 | 95.90 | 79.48 | 95.27 | 94.78 | 81.53 | 94.46 | 93.01 | 82.83 | 94.57 | 92.01 |
| 0-shot | 76.75 | 94.34 | 96.26 | 79.48 | 95.02 | 94.65 | 81.93 | 94.59 | 92.88 | 83.43 | 94.57 | 91.86 |

Table 2: Accuracy on test dataset for each of the 12 few-shot categories using the five-shot BERT-FS$_{r+d}$ model, compared to a zero-shot BERT model which is not fine-tuned on examples from the target domain.

1% over the original model on the corresponding validation data, although these are unstable at higher learning rates. We select a learning rate of 2e-7 for 10 epochs with a batch size of 8 for few-shot learning.

### 5.2 Five-shot classification

We first compare our proposed approach to various other five-shot learning techniques that have been evaluated on the ARSC dataset. The results are reported in Table 1. We find that the simple transfer learning approach outperforms all previous results reported on this dataset apart from Deng et al. (2020), which uses contextual embeddings from a pretrained BERT model to initialize parameters for a meta-learning approach. This result suggests that pretrained models can be easily extended to such few-shot settings, significantly outperforming several meta-learning and metric learning techniques which are trained to generalize from nearly 100k examples. However, the first fine-tuning phase to adapt the model to the review sentiment task appears critical; without this, a model trained only on few-shot data fails to converge and results in a trivial classifier with a mean accuracy of 21.68%.

We also observe that removing the domain string as a prefix does not lead to a large decrease in accuracy. This is surprising as we might expect domain-specific language to be a useful indicator of review sentiment, while some words may imply different sentiment across domains, e.g., “short” may be positive for a movie but negative for a video game. We conjecture that the information of each domain is available to the model through the review text, and that domain-specific indicators of sentiment are less predictive of star ratings than universal lexical signals such as those that express frustration and joy.

### 5.3 Zero-shot classification

In order to further examine the effect of domain on sentiment prediction, we ran additional experiments in a *zero-shot* setting in which we evaluated the models produced by the first fine-tuning phase directly on the few-shot test set, without further fine-tuning. The results for each few-shot class are displayed in Table 2. Surprisingly, this single-phase training approach produces performance similar to the five-shot approach, with minimal changes in accuracy across domains and rating categories. The mean accuracy of the zero-shot model is 89.65%, which is slightly higher than BERT-FS$_{r+d}$, and indicates that the few-shot training hyperparameters optimized on validation domains did not generalize well to the test domains. A closer analysis shows that BERT-FS$_{r+d}$ correctly classifies 7 examples that the zero-shot model misclassifies, but also incorrectly classifies 13 examples that the zero-shot model gets correct. This indicates that it may be beneficial to conduct multiple rounds of hyperparameter search using different surrogate few-shot datasets prior to few-shot learning.

### 5.4 Discussion

The strong performance observed in the zero-shot setting and the modest effect of the domain prefix reveal that domains in the ARSC dataset may not truly be distinct enough to define separate classes. Rich text classifiers such as the ones proposed here may be able to easily generalize over small differences in domain-specific language to make predictions based solely on rating categories. We interpret this outcome as a limitation for the applicability of the ARSC dataset to research on few-shot learning and propose that future work in this area be based on new natural datasets which are demonstrably challenging for models in zero-shot classification settings.

These observations also reinforce the notion that modern transfer learning approaches such as BERT may have taken a great leap forward in terms of domain invariance. The ASRC dataset was proposed in an era when cross-domain sentiment analysis was a significant problem, but by achieving results that are almost state-of-the-art while essentially ignoring domain information, our work suggests that cross-domain knowledge transfer is no longer a major challenge for sentiment analysis tasks.

Finally, our experiments demonstrate that the simple approach of fine-tuning BERT-like models as binary classifiers with class labels supplied as prefixes can yield a powerful baseline in problems involving extreme data sparsity for some labels. With sufficient resources for inference, this technique can be used in a wide variety of multi-class and multi-label classification scenarios and can even generalize effectively to labels that are *never* seen in training. By relying on self-attention to condition the label, this approach avoids the overhead of training separate output layers or adapters (Houlsby et al., 2019) for rare and unseen classes, and the absence of class-specific parameters may further encourage generalization.

## 6 Conclusions & Future Work

We explore the problem of few-shot text classification from the perspective of transfer learning using the popular approach of fine-tuning a pretrained BERT model. Using the well-studied ARSC dataset, we show that the multi-domain sentiment classification problem can be formulated as a single binary classification task where the rating-domain specification from ARSC is encoded within the input. Experimental results reveal that this simple approach outperforms most published results on this dataset, and that a zero-shot model achieves comparable accuracy to a five-shot model. Since we demonstrate that including domain information, which is used to define the classes for this dataset, only yields minor improvements in training a model, the results also suggest that ARSC is no longer an appropriate dataset for studying few-shot learning techniques. We aim to remedy this situation and identify more suitable resources for future research on few-shot and continual text classification.

## References

Trapit Bansal, Rishikesh Jha, and Andrew McCallum. 2019. Learning to few-shot learn across diverse natural language classification tasks. *Computing Research Repository*, arXiv:1911.03863.

Aviad Barzilai and Koby Crammer. 2015. Convex Multi-Task Learning by Clustering. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pages 65–73, San Diego, California, USA.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 440–447, Prague, Czech Republic.

Tom B. Brown, Benjamin Pickman Mann, Nick Ryder, Melanie Subbiah, Jean Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krüger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric J Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher

Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Computing Research Repository*, arXiv:2005.14165.

Shumin Deng, Ningyu Zhang, Zhanlin Sun, Jiaoyan Chen, and Huajun Chen. 2020. When low resource NLP meets unsupervised language model: Meta-pretraining then meta-learning for few-shot text classification. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 13773–13774.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota.

Li Fei-Fei, Rob Fergus, and Pietro Perona. 2006. One-shot learning of object categories. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(4):594–611.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1126–1135.

Ruiying Geng, Binhua Li, Yongbin Li, Xiaodan Zhu, Ping Jian, and Jian Sun. 2019. Induction networks for few-shot text classification. 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 3904–3913, Hong Kong, China.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799, Long Beach, California, USA, 09–15 Jun.

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *Computing Research Repository*, arXiv:1909.05858.

Erik G. Miller, Nicholas E. Matsakis, and Paul A. Viola. 2000. Learning from one example through shared densities on transforms. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 464–471.

Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. 2018. A simple neural attentive meta-learner. In *International Conference on Learning Representations*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *Computing Research Repository*, arXiv:1910.10683.

Sachin Ravi and Hugo Larochelle. 2017. Optimization as a model for few-shot learning. In *International Conference on Learning Representations*.

Victor Garcia Satorras and Joan Bruna. 2018. Few-shot learning with graph neural networks. In *International Conference on Learning Representations*.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems 30*, pages 4077–4087.

Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H. S. Torr, and Timothy M. Hospedales. 2018. Learning to compare: Relation network for few-shot learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1199–1208.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. 2016. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems 29*, pages 3630–3638.

Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni. 2020. Generalizing from a few examples: A survey on few-shot learning. *ACM Comput. Surv.*, 53(3).

Mo Yu, Xiaoxiao Guo, Jinfeng Yi, Shiyu Chang, Saloni Potdar, Yu Cheng, Gerald Tesauro, Haoyu Wang, and Bowen Zhou. 2018. Diverse few-shot text classification with multiple metrics. 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 1206–1215, New Orleans, Louisiana.