On Dataset Transferability in Active Learning for Transformers

Fran Jelenić **Josip Jukić** **Nina Drobac** **Jan Šnajder**
University of Zagreb, Faculty of Electrical Engineering and Computing, Croatia
Text Analysis and Knowledge Engineering Lab
{fran.jelenic, josip.jukic, nina.drobac, jan.snajder}@fer.hr

Abstract

Active learning (AL) aims to reduce labeling costs by querying the examples most beneficial for model learning. While the effectiveness of AL for fine-tuning transformer-based pre-trained language models (PLMs) has been demonstrated, it is less clear to what extent the AL gains obtained with one model transfer to others. We consider the problem of transferability of actively acquired datasets in text classification and investigate whether AL gains persist when a dataset built using AL coupled with a specific PLM is used to train a different PLM. We link the AL dataset transferability to the similarity of instances queried by the different PLMs and show that AL methods with similar acquisition sequences produce highly transferable datasets regardless of the models used. Additionally, we show that the similarity of acquisition sequences is influenced more by the choice of the AL method than the choice of the model.

1 Introduction

Pre-trained language models (PLMs) – large overparameterized models based on the transformer architecture (Vaswani et al., 2017) and trained on large corpora – are the leading paradigm in modern NLP, yielding state-of-the-art results on a wide range of NLP tasks. However, large models require large amounts of data. *Active learning* (**AL**; Settles, 2009) addresses the data bottleneck problem by improving data labeling efficiency. It employs human-in-the-loop labeling with the model iteratively selecting data points most informative for labeling. Recent work has demonstrated the effectiveness of AL for fine-tuning PLMs (Dor et al., 2020; Grießhaber et al., 2020; Margatina et al., 2022; Yuan et al., 2020; Shelmanov et al., 2021).

While AL may considerably reduce model development costs, it also potentially limits the scope of use of the actively acquired datasets. Since data sampling in AL is guided by the inductive bias of the acquisition model, the dataset will typically not represent the original population's distribution (Attenberg and Provost, 2011). This is troublesome if one wishes to use the actively acquired dataset to train a different model (*consumer model*) from the one used for AL (*acquisition model*). If the two models' inductive biases differ, the AL gains can cancel or even revert: the consumer model may perform worse when trained on the actively acquired dataset than on a randomly sampled one. However, the robustness of the actively acquired dataset to the choice of the consumer model is obviously highly desirable, as the acquisition model may become unavailable or dated. The latter is common in NLP, where new and better models are being developed faster than new datasets. However, most AL studies use the same acquisition and consumer models, and dataset transferability is seldom mentioned in AL literature. A notable exception is the work of Lowell et al. (2018), who showed the unreliability of dataset transfer on standard NLP tasks.

In this work, we examine the problem of AL dataset transferability for transformer-based PLMs and conduct a preliminary empirical study on text classification datasets. We first probe whether AL gains persist between different transformer-based PLMs, considering several AL methods and datasets. Observing that on most datasets, the transfer works in some cases but fails in others, we investigate the mechanisms underlying transferability. We hypothesize a link between AL dataset transferability and how the acquisition and consumer models sample instances. To probe this, we introduce *acquisition sequence mismatch* (ASM) to characterize to what extent the two models differ in how they sample instances throughout AL iterations. We investigate how ASM affects dataset transferability and how ASM is affected by other AL variables. We show that, while it is generally reasonable to transfer actively acquired datasets between transformer-based PLMs, AL methods that

Findings of the Association for Computational Linguistics: ACL 2023, pages 2282–2295
July 9-14, 2023

retain low ASM produce more transferable datasets. We also show that the choice of the AL method affects ASM more than the choice of models.

To summarize our contributions: we (1) conduct an empirical study on the transferability of actively acquired datasets between transformer-based PLMs, (2) propose a measure to quantify the mismatch in the acquisition sequences of AL models and link this to dataset transferability, and (3) analyze what design choices affect this mismatch. We provide code for the experiments[1] with the hope that our results will encourage NLP practitioners to use AL when fine-tuning PLMs and motivate further research into the AL dataset's transferability.

2 Related Work

Although AL has been extensively studied for shallow and standard neural models (without pretraining), research on combining AL and PLMs lags behind. The initial studies showed promise, with AL methods outperforming random sampling for text classification (Dor et al., 2020; Grießhaber et al., 2020). The field is gradually gaining traction with studies demonstrating AL effectiveness even with simple uncertainty-based methods (Gonsior et al., 2022; Schröder et al., 2022). Moreover, PLMs open up new possibilities, such as complementing AL with model adaptation using unlabeled data (Yuan et al., 2020; Margatina et al., 2022).

While there is much research on AL for standard scenarios where the acquisition and consumer models are the same, there is little research on AL dataset transfer. Prabhu et al. (2019) demonstrated that combining uncertainty AL strategies with deep models produces sampled datasets with good sampling properties that have a large overlap with support vectors of SVM trained on the entire dataset. Likewise, Farquhar et al. (2021) showed that deep neural models benefit from the sample bias induced by the acquisition model (the opposite is true for shallow models). However, the jury is still out on the effects of sample bias on the consumer model. The most prominent empirical study on AL transfer with neural models (Lowell et al., 2018) predates PLMs. Tsvigun et al. (2022) focused on alleviating the effects of acquisition-consumer mismatch in PLMs by using lightweight distilled models for acquisition and larger versions of the models as consumer models. Even though the study focuses on improving the transferability of actively acquired datasets, the reasons behind the successful transfer are yet to be explored. An older study of AL dataset transferability for text classification and shallow models by Tomanek and Morik (2011) showed that transfer works in most cases but that neither sample nor model similarity explains transferability. Our study explores these characteristics for acquisition-consumer pairings of different PLMs.

3 Experimental Setup

Our study used four datasets, three models, and three AL methods (cf. Appendix B for details). The datasets we used are Subjectivity (**SUBJ**; Pang and Lee, 2004), CoLA (**COLA**; Warstadt et al., 2018), AG-News (**AGN**; Zhang et al., 2015), and TREC (**TREC**; Li and Roth, 2002)). The three transformer models we used are BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), and ELECTRA (Clark et al., 2020). The AL methods we considered are entropy (**ENT**; Settles, 2009), core-set (**CS**; Sener and Savarese, 2017), and BADGE (**BA**; Ash et al., 2019)). This gives 108 AL configurations (72 transfer and 36 no-transfer configurations). Furthermore, we ran each configuration with 20 different warm-start sets to account for stochasticity. The AL acquisition was simulated until the budget of 1500 labeled data points was exhausted (model performance for all datasets reached a plateau), labeling 50 data points per step.

We assessed dataset transferability using the difference in the area under the F_1 curve of the model trained on the actively acquired dataset and the same model trained on a randomly sampled dataset (ΔAUC). We deem the AL dataset transfer successful if ΔAUC is not significantly less than zero and unsuccessful otherwise. We chose ΔAUC to make the notion of transferability independent of when the AL acquisition terminates. On the other hand, as terminating the AL after acquiring too few labeled data is unrealistic, we also report ΔAUC_{10}, which is ΔAUC calculated with an offset of 10 iterations (500 labeled instances) of the AL loop. Comparing ΔAUC_{10} to ΔAUC provides insights into how transferability changes through time.

4 Results

4.1 Dataset transferability

We grouped the 108 AL configurations into three groups based on the sign of the mean ΔAUC value and the p-value of the difference between AUC

[1] https://github.com/fjelenic/al-transfer

	Δ^-	Δ^0	Δ^+	Δ^-_{10}	Δ^0_{10}	Δ^+_{10}	Σ
SUBJ	0	0	18	0	0	18	18
COLA	2	8	8	2	7	9	18
AGN	7	4	7	3	2	13	18
TREC	8	3	7	0	2	16	18
R→B	2	2	8	0	1	11	12
E→B	2	2	8	0	2	10	12
B→R	2	4	6	0	1	11	12
E→R	2	4	6	1	2	9	12
B→E	5	1	6	2	2	8	12
R→E	4	2	6	2	3	7	12
ENT	11	3	10	3	2	19	24
CS	4	10	10	2	6	16	24
BA	2	2	20	0	3	21	24
Σ	17	15	40	5	11	56	

Table 1: Breakdown of datasets, acquisition→consumer model pairs (denoted by initial letters), and AL methods by transferability: negative ($-$), neutral (0), and positive ($+$) transfer. ΔAUC is shown as Δ.

scores of transfer and random sampling:[2] negative ($\Delta\text{AUC} < 0$ and p$<$.05), neutral (p$\geq$.05), and positive ($\Delta\text{AUC} \geq 0$ and p$<$.05) transfer. The no-transfer AL configurations (where the acquisition and consumer models are the same) are generally successful (25 positive, 9 neutral, and 2 negative configurations as per ΔAUC; 33 positive, 2 neutral, and 1 negative configuration as per ΔAUC_{10}). The grouping of the remaining 72 configurations with AL dataset transfer is given in Table 1. We observe that the dataset, the acquisition-consumer model pairing, and the AL method all affect transfer success.

Evidently, transferability differs across datasets: the transfer is always positive on SUBJ (which is the simplest task we considered in terms of the number of labels, the balance of classes, and the MDL task complexity measure; cf. Appendix B), while most neutral transfers occur on COLA. A more interesting picture emerges from the different acquisition-consumer model pairings and AL methods. Most negative transfers are transfers to ELECTRA, while most neutral transfers are those to RoBERTa (perhaps due to it being optimized for robustness). On the other hand, transfer to BERT is positive in most cases, perhaps because BERT's pre-training regime is most similar to that of the other two models. Among the AL methods, entropy mostly makes the transfer negative, most neutral transfers occur with core-set, and BADGE is the best choice for ensuring positive transferability. However, when looking at the later steps of the AL loop, differences between entropy and BADGE vanish, while the core-set lags slightly behind. Thus, ΔAUC tends to increase throughout the AL process, suggesting that increasing the amount of sampled data lowers the risk of unsuccessful transfer (cf. Appendix C for additional F_1 scores analysis).

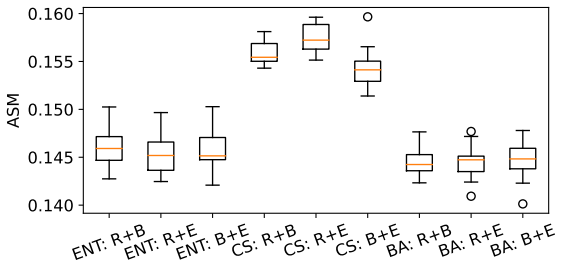
[2]We used either the paired t-test or Wilcoxon signed-rank test, depending on the results of Lilliefors' test for normality.

4.2 Acquisition sequence mismatch

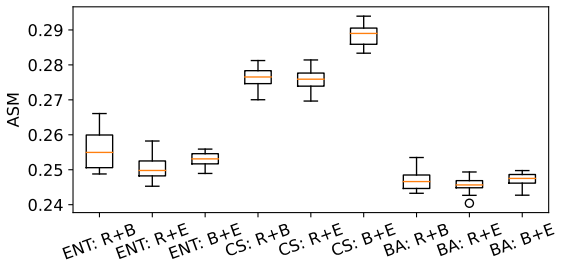
We hypothesize there is a link between dataset transferability and the sequence in which data points are acquired for labeling by AL. In particular, we posit that dataset transferability will be successful when the acquisition sequence of the acquisition model does not differ from what the acquisition sequence of a consumer model would be if that model had access to the original dataset. We introduce the *acquisition sequence mismatch* (ASM) to measure the differences in acquisition sequences. To compute the ASM between two acquisition sequences, we pair the corresponding batches of the two sequences and average their pairwise differences. To measure the difference between a pair of batches, we take the average of the distances of best-matched examples between the batches. To account for the fact that AL methods may choose numerically different yet semantically similar data points, we measure the similarity of acquired instances in representation space. We use GloVe embeddings (Pennington et al., 2014) as a common representation space independent of the choice of acquisition and consumer models and compute the cosine distance between averaged word embeddings. Lastly, we use the Hungarian algorithm (Kuhn, 1955) to construct a bipartite graph between two batches with distance-weighted edges to find the best-matching examples. Formally, we define ASM as follows:

$$\frac{1}{T}\sum_{t=1}^{T}\frac{1}{|B_t|}\min_{S(B^t_A),S(B^t_B)}\left(\sum_{i=1}^{|B_t|}d(x^i_A,x^i_B)\right) \quad (1)$$

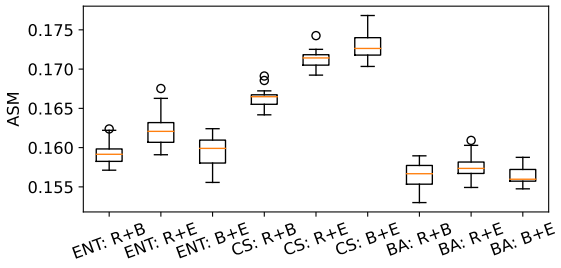
where T is the length of the sequence (the number of steps of the AL loop), $S(B^t)$ is the set of all of the permutations of instances in the selected batch at step t, and $d(x^i_A, x^i_B)$ is the cosine distance between instance representations from sequences A and B for a batch at position i of a given batch permutation. Intuitively, ASM assumes that both batches cater to the same informational need of the



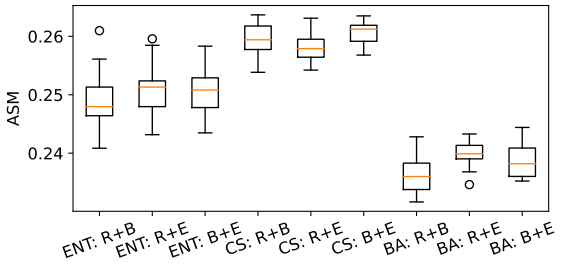
(a) Subjectivity



(b) CoLA



(c) AG-News



(d) TREC

Figure 1: Distributions of ASM values for combinations of AL methods and acquisition+consumer model pairs (denoted by initial letters).

model, so it calculates how much the instances that should carry out the same role in the batch differ.

Given a dataset, we hypothesize ASM may be affected by both the choice of the models and the choice of the AL method. Figure 1 shows that the distributions of ASM values are more alike when grouped by the AL methods than when grouped by the model pairings. To verify this observation, we conducted two Kruskal-Wallis H-tests for each dataset: in the first, populations were determined by the AL method, and we concluded that there was a significant difference in ASM ($p<.05$); in the second, the populations were determined by the model pairing, and there was no significant difference in ASM ($p>.05$). This suggests that the choice of AL method affects ASM more than the choice of acquisition-consumer model pairing.

4.3 Acquisition mismatch analysis

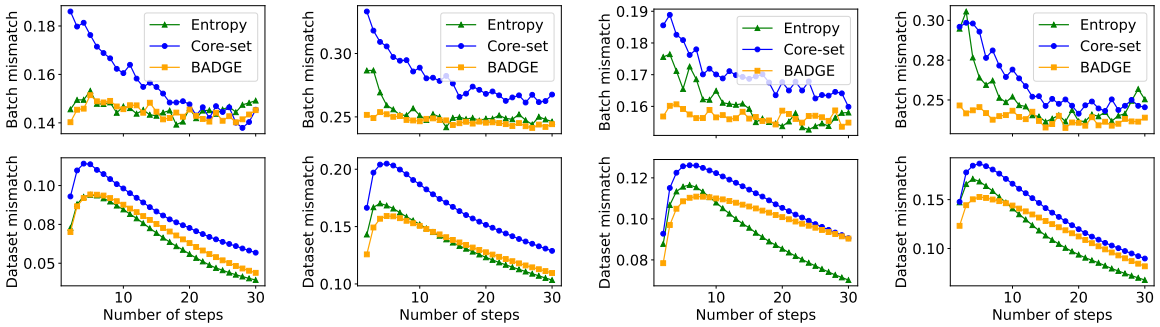
We found a statistically significant negative correlation between ΔAUC and ASM for each dataset.[3] This supports our hypothesis that the lower the mismatch between acquisition sequences of the two models, the higher the transferability of a dataset from one model to the other. Besides ASM, we use another measure for analyzing dataset transferability: the difference between the dataset acquired with AL using the acquisition model and the dataset acquired with AL using the consumer model. We call this measure the *acquired dataset mismatch* (ADM). Essentially, ADM computes the mismatch between samples similarly to ASM but between entire datasets obtained after the last sampling step.

Above we showed that the choice of the AL method affects the ASM. Figure 2 shows that BADGE gives smaller ASM than the other two methods, whereas core-set gives larger ASM than the other two methods.[4] However, the intriguing effect emerges when comparing the difference in batches through time and differences in the entire acquired datasets through time. In the early steps, BADGE gives the highest similarity of acquired datasets among the considered methods, which leads to it having the lowest ASM. However, in later steps, entropy dominates the similarity of acquired datasets.[5] It seems as if entropy acquired similar datasets for different models by taking those models through different sequences of the population distribution. This effect is seen in Table 1, where entropy is the worst method when using ΔAUC to measure transfer success while managing to parry BADGE when using ΔAUC_{10}. The difference in transferability between entropy and BADGE completely vanishes when looking at the last step of the AL loop (cf. Appendix, Table 3).

[3] Spearman correlation coefficients are -0.11 for SUBJ, -0.19 for COLA, -0.27 for AGN, and -0.38 for TREC, all significant with $p<.05$.

[4] Verified using three one-sided Wilcoxon signed-rank tests with $p<.05$ corrected for FWER.

[5] Verified using three one-sided Wilcoxon signed-rank tests on ADM with $p<.05$ corrected for FWER.



(a) Subjectivity (b) CoLA (c) AG-News (d) TREC

Figure 2: The mismatch between acquired batches (top) and ADM at each step of the AL loop (bottom) for different AL methods.

It is clear that entropy can produce transferable datasets, but it requires more time to do so.

We speculate that the effect of BADGE having the lowest ASM yet entropy achieving the lowest ADM could emerge due to the interaction between the AL method and the model's decision boundary. Namely, uncertainty AL methods sample data points on the decision boundary with high overlap with support vectors of the SVM trained on the whole dataset, as pointed out by Prabhu et al. (2019). Since BADGE combines uncertainty and diversity, i.e., it samples data points the model is uncertain about for diverse reasons, it samples along the entire decision boundary at each step, and since decision boundaries of the models are roughly the same, so are the sampled data points. Entropy, on the other hand, relies solely on uncertainty. Due to its greedy nature, entropy tends to sample similar points because if one data point has high uncertainty, data points similar to it are also going to have high uncertainty (Zhdanov, 2019). This may manifest as sampling local patches of space on the decision boundary. Therefore, entropy may take more time to define the boundary than BADGE because it is forming the boundary from patches of space with the highest uncertainty at a given AL step rather than holistically sampling along the boundary at each step. Since the shape of the decision boundary is more similar between different models than the local interactions along the boundary, entropy has a higher batch mismatch in the early steps. However, once more data is labeled and the boundary becomes stable, both entropy and BADGE start to have a low batch mismatch, as seen in Figure 2. Since entropy is deterministic and never strays from the decision boundary, it ends up having a lower ADM than BADGE. Lastly, we believe that the core-set method has the highest ASM and ADM because it selects data based on diversity in the model's representation space, which is more model-specific and shares fewer properties between different models than the decision boundary. Further exploring the described interaction is a compelling direction for future work.

It may be that AL methods with different acquisition sequences end up acquiring a similar dataset and have high transferability, as in the case of entropy, an uncertainty-based acquisition function. It is also possible that acquired datasets differ between models but that the transfer remains successful because it taps into some other essential aspect of a transferable dataset, as is the case with core-set, a diversity-based acquisition function. However, the best strategy to ensure dataset transferability appears to be a mixture of uncertainty and diversity, as provided by BADGE. This appears to minimize ASM between models, making datasets transferable regardless of the number of AL steps.

5 Conclusion

We presented an empirical study on the transferability of actively acquired text classification datasets for transformer-based PLMs. Our results indicate no significant risk in transferring datasets, especially for larger amounts of data. We also showed that transfer is largely successful when preserving the sequence and similarity of acquired instances between the models, which is what methods combining uncertainty and diversity acquisition functions seem to do. Transferability appears to differ considerably across datasets, so future work should examine what dataset characteristics are predictive of transfer success.

Limitations

Our study revealed considerable differences in transferability and other measures we considered across different datasets. Nonetheless, the study focused on the differences in transferability arising from the choice of the models and the AL methods rather than the dataset. To eliminate confounding due to datasets, we grouped the results by datasets and analyzed each group separately. Despite this, the scope of our results is limited by the fact that all datasets used are in English and possibly contain their own biases.

Even though we showed that it could still be useful to transfer actively acquired datasets between transformer-based PLMs, it is important to keep in mind that actively acquired datasets are not representative of the original data distribution due to the sampling bias introduced by active learning.

Acknowledgments

This research was supported by the AIDWAS KK.01.2.1.02.0285 grant. We thank the anonymous reviewers for their insightful comments and suggestions.

References

Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. 2019. Deep batch active learning by diverse, uncertain gradient lower bounds. *arXiv preprint arXiv:1906.03671*.

Josh Attenberg and Foster Provost. 2011. Inactive learning? difficulties employing active learning in practice. *ACM SIGKDD Explorations Newsletter*, 12(2):36–41.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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.

Sebastian Farquhar, Yarin Gal, and Tom Rainforth. 2021. On statistical bias in active learning: How and when to fix it. *arXiv preprint arXiv:2101.11665*.

Julius Gonsior, Christian Falkenberg, Silvio Magino, Anja Reusch, Maik Thiele, and Wolfgang Lehner. 2022. To softmax, or not to softmax: that is the question when applying active learning for transformer models. *arXiv preprint arXiv:2210.03005*.

Daniel Grießhaber, Johannes Maucher, and Ngoc Thang Vu. 2020. Fine-tuning BERT for low-resource natural language understanding via active learning. *arXiv preprint arXiv:2012.02462*.

Harold W Kuhn. 1955. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

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 preprint arXiv:1907.11692*.

David Lowell, Zachary C Lipton, and Byron C Wallace. 2018. Practical obstacles to deploying active learning. *arXiv preprint arXiv:1807.04801*.

Katerina Margatina, Loïc Barrault, and Nikolaos Aletras. 2022. On the importance of effectively adapting pretrained language models for active learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 825–836.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the ACL*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. Rissanen data analysis: Examining dataset characteristics via description length. In *International Conference on Machine Learning*, pages 8500–8513. PMLR.

Ameya Prabhu, Charles Dognin, and Maneesh Singh. 2019. Sampling bias in deep active classification: An empirical study. 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 4058–4068.

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.

Ozan Sener and Silvio Savarese. 2017. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*.

Burr Settles. 2009. Active learning literature survey. Computer sciences technical report.

Artem Shelmanov, Dmitri Puzyrev, Lyubov Kupriyanova, Denis Belyakov, Daniil Larionov, Nikita Khromov, Olga Kozlova, Ekaterina Artemova, Dmitry V Dylov, and Alexander Panchenko. 2021. Active learning for sequence tagging with deep pre-trained models and bayesian uncertainty estimates. *arXiv preprint arXiv:2101.08133*.

Katrin Tomanek and Katherina Morik. 2011. Inspecting sample reusability for active learning. In *Active Learning and Experimental Design workshop In conjunction with AISTATS 2010*, pages 169–181. JMLR Workshop and Conference Proceedings.

Akim Tsvigun, Artem Shelmanov, Gleb Kuzmin, Leonid Sanochkin, Daniil Larionov, Gleb Gusev, Manvel Avetisian, and Leonid Zhukov. 2022. Towards computationally feasible deep active learning. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1198–1218.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2018. Neural network acceptability judgments. *arXiv preprint arXiv:1805.12471*.

Michelle Yuan, Hsuan-Tien Lin, and Jordan Boyd-Graber. 2020. Cold-start active learning through self-supervised language modeling. *arXiv preprint arXiv:2010.09535*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

Fedor Zhdanov. 2019. Diverse mini-batch active learning. *arXiv preprint arXiv:1901.05954*.

	Train	Test	# Labels	NLE	MDL
SUBJ	8000	2000	2	1.00	0.30
COLA	8551	1043	2	0.88	1.00
AGN	20000*	7600	4	1.00	0.56
TREC	5452	500	6	0.92	0.34

Table 2: Dataset statistics. We report train and test set sizes, number of labels, normalized label entropy (information entropy of label distribution normalized by the entropy of uniform distribution with the same number of variables), and MDL normalized by the MDL of the dataset with the largest value (COLA). Train set of AGN was subsampled from the original train set of 120000 instances.

A Reproducibility

We conducted our experiments on 4× AMD Ryzen Threadripper 3970X 32-Core Processors and 4× NVIDIA GeForce RTX 3090 GPUs with 24GB of RAM, which took roughly one week. We used PyTorch version 1.12.1, Transformers version 4.21.3, and CUDA 11.4.

B Experimental design choices

B.1 Datasets

The datasets used in this paper are standard benchmarks in NLP for text classification. We chose these datasets to represent different attributes: the number of labels (binary or multi-class classification) and the balancing of the labels (balanced and imbalanced classes). The diversity of the dataset characteristics can give an insight into the impact of these attributes on dataset transferability. We present dataset statistics in Table 2. There we also show *minimum description length* (MDL) (Perez et al., 2021) of each dataset, which can be interpreted as the complexity of the task.

Subjectivity: Movie-review data with reviews labeled as either subjective or objective. This is a balanced dataset with binary labels.

CoLA: The Corpus of Linguistic Acceptability is a dataset containing sentences labeled as grammatical or not. This is an imbalanced dataset with binary labels.

AG-News: Corpus of news articles annotated by the article's topic (World, Sports, Business, Sci/Tech). The dataset was created by subsampling the corpus to the size of 20,000 examples. This is a balanced dataset with four classes.

TREC: The dataset contains questions labeled with the type of subject of the question. This is an imbalanced dataset with six classes.

B.2 Models

We picked the models that share the common architecture; they are all transformer-based PLMs but differ in pre-training data and pre-training objectives. This choice of models enables us to analyze the impact of different pre-training design choices on dataset transferability. All models were trained using ADAM optimizer with a learning rate of $2 \cdot 10^{-5}$ and batch size of 64 for five epochs for both acquisition and evaluation phases.

BERT: One of the first and most popular transformer-based pre-trained language models. The model was pre-trained using a generative masked language modeling objective. This model has 12 layers, a hidden state size of 768, and 12 heads with 110M parameters in total.

RoBERTa: A model with the same architecture and pre-training objective as BERT but trained on more data and with optimized hyperparameters to make the model more robust. This model has 12 layers, a hidden state size of 768, and 12 heads with 125M parameters in total.

ELECTRA: It uses the same architecture and pre-training data as BERT but with discriminative instead of generative pre-training objectives. Instead of masking some tokens in text and having to guess the identity of masked tokens as BERT does, the generative pre-training objective corrupts some tokens by replacing them with plausible alternatives, and then the model has to decide for each token whether it is the original token or the replaced one. This model has 12 layers, a hidden state size of 768, and 12 heads with 110M parameters in total.

B.3 AL methods

AL methods used to select the most informative data points are divided into two types of heuristics: uncertainty and diversity. Methods using uncertainty as a heuristic select data based on some measure of the model's uncertainty. The intuition behind the uncertainty methods is that the more uncertain the model is about a data point, the more it can learn from knowing its label. In comparison, diversity-based methods try to represent the input space (which is not always the same as the input population) as accurately as possible with as few data points as possible. AL methods can combine those two heuristics to select a group of data points the model is uncertain about for different reasons.

The choice of the AL methods used in this experiment was motivated by the type of heuristic

	F_1^-	F_1^0	F_1^+	
SUBJ	0	1	17	18
COLA	3	11	4	18
AGN	1	4	13	18
TREC	0	2	16	18
R→B	0	3	9	12
E→B	0	4	8	12
B→R	1	1	10	12
E→R	1	2	9	12
B→E	1	3	8	12
R→E	1	5	6	12
ENT	1	5	18	24
CS	2	8	14	24
BA	1	5	18	24
	4	18	50	

Table 3: Breakdown of datasets, acquisition→consumer model pairs (denoted by initial letters), and AL methods by transferability measured via F_1 score at the end of the AL loop: negative ($-$), neutral (0), positive ($+$) transfer.

(uncertainty vs. diversity) they used for sampling. These methods allow us to analyze the impact of the choice of heuristic on the success of dataset transfer in AL.

Entropy: An uncertainty-based method that selects data points with maximal information entropy of their posterior class distribution.

Core-set: This diversity-based method selects data points that best cover the representation space.

BADGE: A method that combines uncertainty and diversity by using k-MEANS++ algorithm on the would-be gradients of the models' last layer for the data points if their most probable labels were their actual labels.

C Experiment runs

This section presents more results from our experiment to complement the already presented results. Table 3 shows transferability for different combinations in the fashion of Table 1. However, instead of measuring transferability with ΔAUC this table uses the F_1 score at the end of the AL loop (1500 labeled instances). To illustrate the success of regular AL (without the transfer), we present Table 4. That table shows the same information as Table 1 and Table 3 but for situations where acquisition and consumer models are the same. Lastly, we present the learning curves of all of the runs of the experiment in Figure 3 for Subjectivity, Figure 4 for CoLA, Figure 5 for AG-News, and Figure 6 for TREC dataset.

	AUC^{-}	AUC^{0}	AUC^{+}	AUC^{-}_{10}	AUC^{0}_{10}	AUC^{+}_{10}	F_1^{-}	F_1^{0}	F_1^{+}	
SUBJ	0	0	9	0	0	9	0	0	9	9
COLA	0	3	6	1	2	6	1	4	4	9
AGN	1	3	5	0	0	9	0	0	9	9
TREC	1	3	5	0	0	9	0	0	9	9
BERT	1	2	9	0	0	12	0	0	12	12
RoBERTa	0	4	8	1	2	9	0	3	9	12
ELECTRA	1	3	8	0	0	12	1	1	10	12
ENT	1	4	7	0	1	11	0	1	11	12
CS	1	4	7	1	0	11	0	2	10	12
BA	0	1	11	0	1	11	1	1	10	12
	2	9	25	1	2	33	1	4	31	

Table 4: Breakdown of datasets, models, and AL methods in the groups based on the performance of regular AL: negative ($-$), neutral (0), positive ($+$) AL.

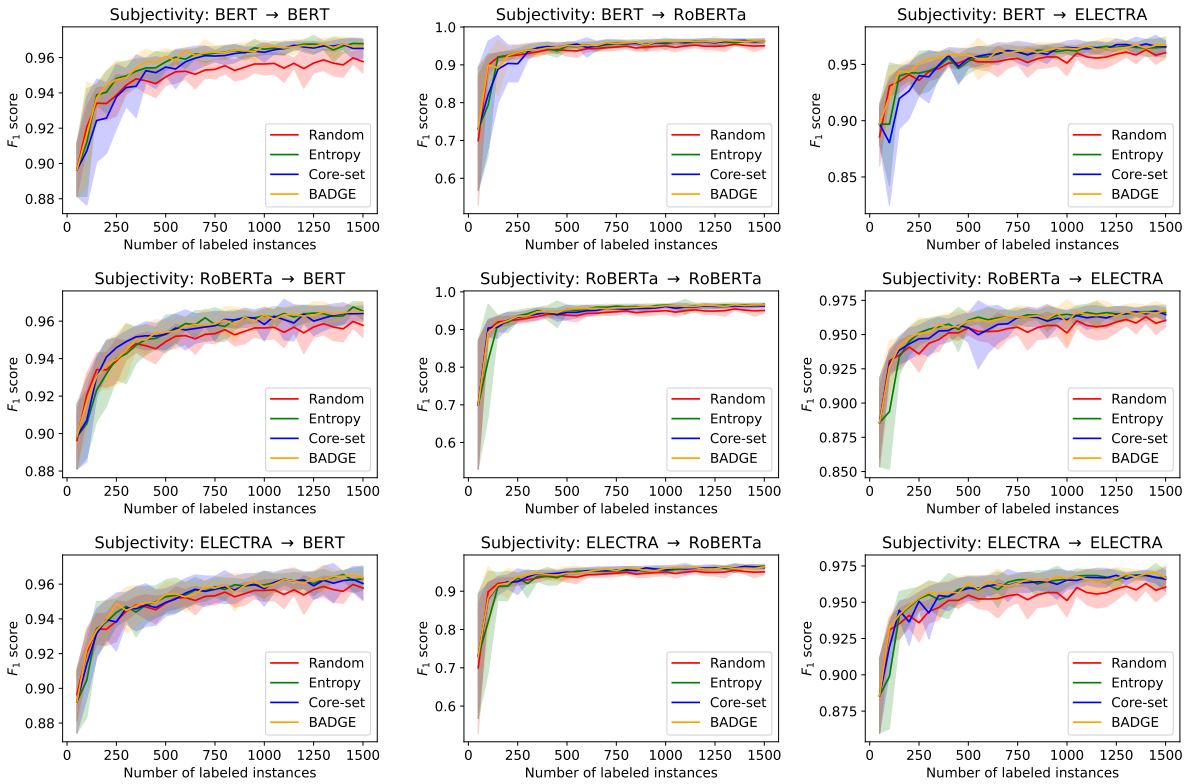


Figure 3: Learning curves for the Subjectivity dataset. The figure shows the mean F_1 score of 20 runs with confidence intervals of $\pm$ standard deviation.

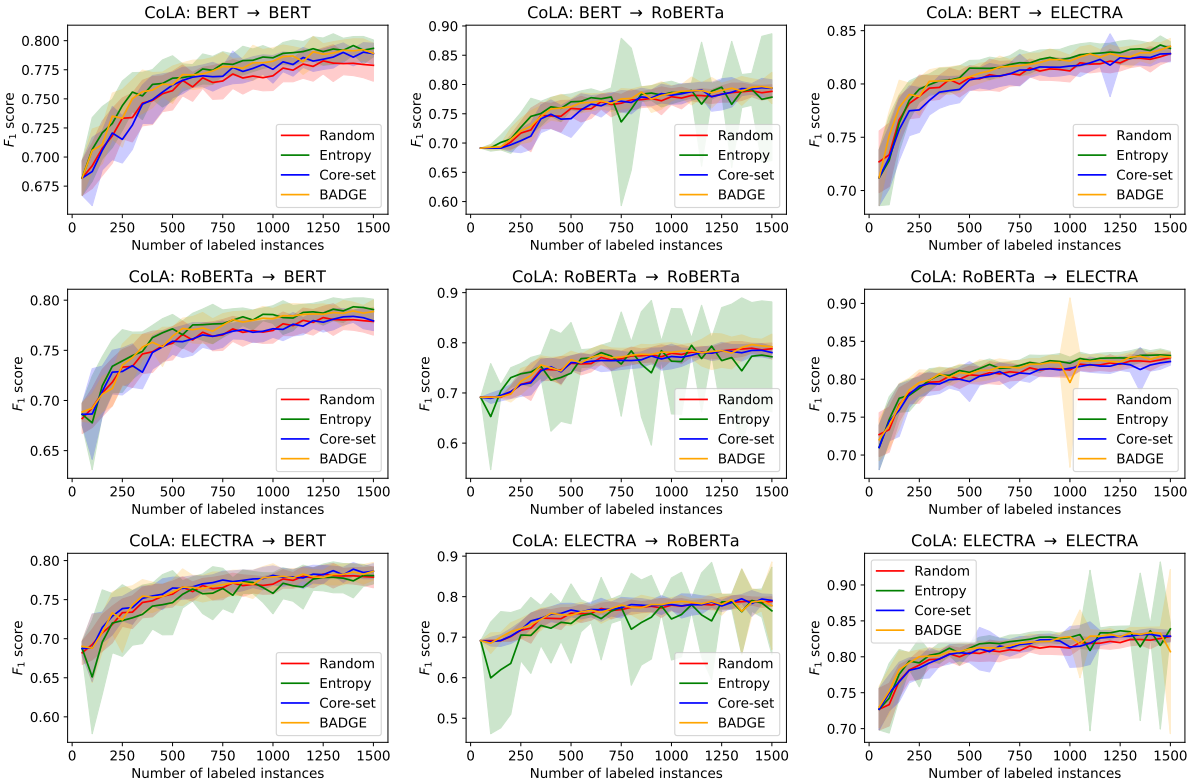


Figure 4: Learning curves for the CoLA dataset. The figure shows the mean F_1 score of 20 runs with confidence intervals of $\pm$ standard deviation. F_1 curves for RoBERTa as consumer model with entropy as AL method have high variance because entropy tends to favor minority class heavily, and the model starts to classify with minority class more often than it should, so the F_1 on the test set drastically drops. These drops happen one to two times per seed during the AL loop before the method balances out the labels again.

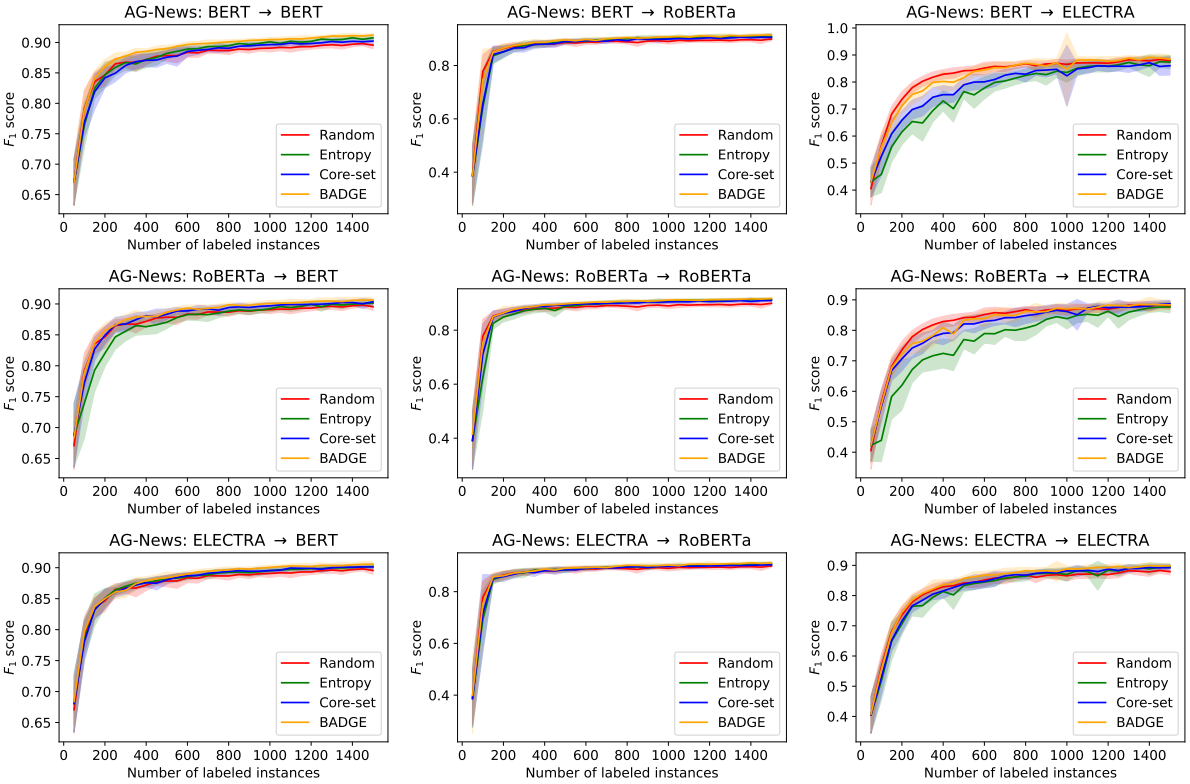


Figure 5: Learning curves for the AG-News dataset. The figure shows the mean F_1 score of 20 runs with confidence intervals of $\pm$ standard deviation.

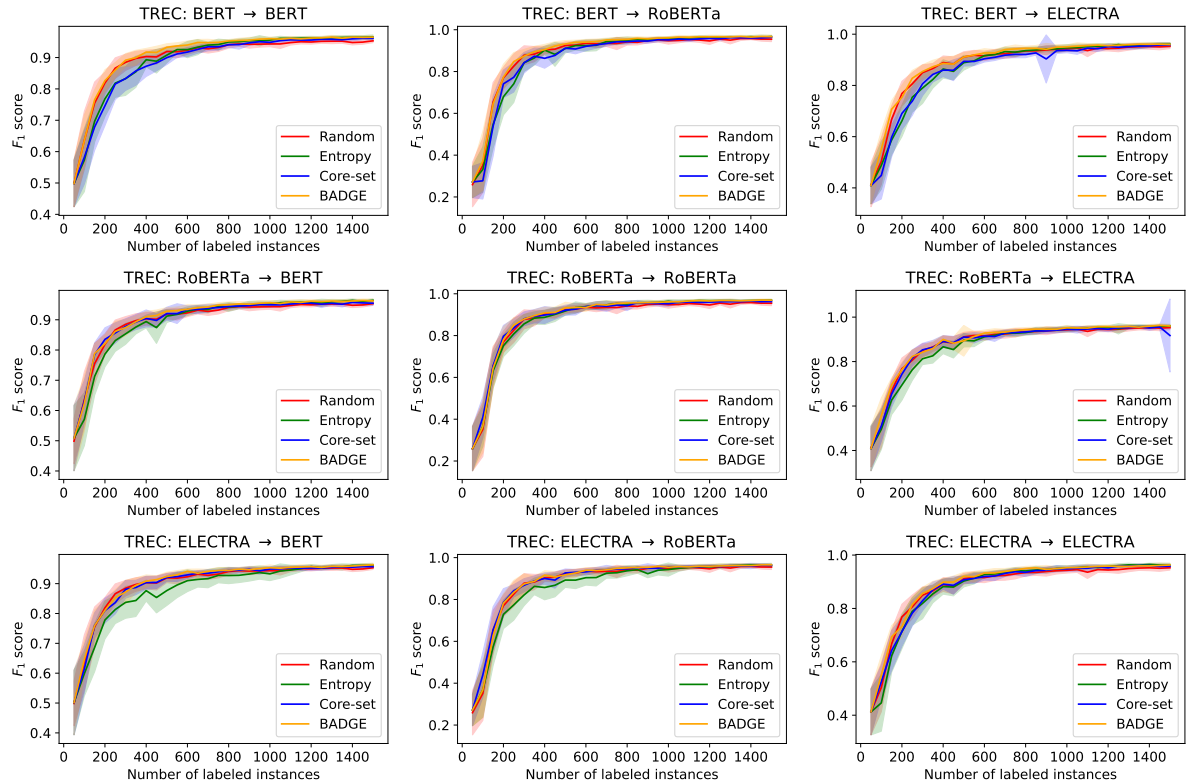


Figure 6: Learning curves for the TREC dataset. The figure shows the mean F_1 score of 20 runs with confidence intervals of $\pm$ standard deviation.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
(Not numbered) Limitations

☑ A2. Did you discuss any potential risks of your work?
(Not numbered) Limitations

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract, 1 Introduction

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

3 Experimental Setup

☑ B1. Did you cite the creators of artifacts you used?
3 Experimental Setup

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
All are standard NLP datasets and models.

☒ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
All are standard NLP datasets and models.

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
All are standard NLP datasets and models.

☒ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
All are standard NLP datasets and models.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
(Appendix) B Experimental design choices

C ☑ Did you run computational experiments?

3 Experimental Setup, 4 Results

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
No response.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
3 Experimental Setup, (Appendix) B Experimental design choices

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
4 Results

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
(Appendix) A Reproducibility, (Appendix) B Experimental design choices

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.