Exploring Transfer Learning Performance in NLP: A Cross-Dataset Generalization Study

Hildah Ngondoki, Peng Zhao, Qiong Zhang

Abstract

Transfer learning has emerged as a cornerstone of modern natural language processing (NLP), particularly with transformer-based architectures like BERT and its variants. Despite substantial progress, critical questions remain regarding how best to transfer knowledge across domains, the influence of training data sequences, and the effectiveness of different BERT models in transfer scenarios. In this paper, we investigate the transferability and generalization capability of RoBERTa, DistilBERT, and DeBERTa across related and unrelated domains using the Amazon Reviews dataset. We design three pipelines to analyze domain similarity, sequencing, dataset sizes, and fine-tuning strategies. Our results demonstrate the importance of sequencing, domain similarity, and hyperparameter tuning in optimizing cross-domain performance. DeBERTa shows superior cross-domain transfer, while RoBERTa excels in low-resource fine-tuning. These findings provide practical insights for leveraging transfer learning in real-world NLP applications.

1 Introduction

In recent years, transformer-based models, particularly the BERT family (Devlin et al., 2019), have revolutionized the field of natural language processing (NLP). These models are pre-trained on large corpora and can be fine-tuned on specific tasks, making them an ideal candidate for transfer learning. Transfer learning involves transferring knowledge gained from one task to another, typically by fine-tuning a pre-trained model on a smaller, task-specific dataset(Gholizade et al., 2025). A subset of this is cross-domain transfer learning, in which knowledge from training models in one domain could help models perform better in another domain (Yesmin, 2024). Despite its transformative success, critical gaps persist in understanding the optimal application of transfer learning. Our work addresses these challenges by systematically evaluating model performance across leading BERT variants, the role of dataset relatedness in knowledge transfer, hyperparameter -tuning strategies for maximal efficiency, the minimal data threshold for effective adaptation, and the influence of training data sequence in continual learning. This paper aims to answer the following research questions:

1) Which BERT family model works best for transfer learning in NLP classification tasks?
2) How does model performance vary when trained on related versus unrelated datasets?
3) What are the optimal fine-tuning strategies for transfer learning?
4) How much data is needed to achieve effective transfer learning effects,that is, 1,000, 5,000 records or 10,000 records?
5) Does the sequence of data training in continuous learning matter?

We explore these questions through a series of experiments targeting binary classification tasks for sentiment analysis by comparing the performance of three different BERT models-RoBERTa, DistilBERT and DeBERTa. We use 3 datasets, two of which are in similar domains, and one in a different domain. We explore varied hyperparameter-tuning strategies, use different sequencing of the datasets to detect transfer learning, and test different training data sizes for effectiveness of transfer learning. To measure each model's capability of generalization, we use accuracy, and F1-score metrics to evaluate the proportion of correctly predicted instances and the validation loss to determine learning efficiency. To our knowledge, this is the first comparative study that systematically combines dataset sequence variation, varying training sizes, intermediate training and three major BERT variants across controlled sentiment tasks. By understanding these aspects, we aim to provide a guide and alternative approaches for practitioners seeking to apply transfer learning to NLP tasks.

2 Related Work

There have been several studies that have shown evidence of positive transfer learning especially for low resource data sets. Yosinski,J et al. (2024) found that initializing with transferred features can improve generalization performance, whilst noting that there was optimization difficulties related to splitting networks in the middle of fragilely co-adapted layers. Yada et al(2020) experimented on intermediate tasks with complex reasoning for Finetuning ROBERTa and noted that it was difficult to draw a conclusion on other factors that drove the positive transfer. Kouw, W.M. (2018) found the complexity of general cases of domain shifts due to multiple factors changing at the same time cannot always be uniquely identified. Vu et al. (2020) also conducted an in-depth analysis of transferability across various NLP tasks and found that transfer learning works best when the pre-trained model is fine-tuned on a task similar to the one it was originally trained on. This aligns with the intuition that related datasets tend to yield better performance when models are adapted to them.

However, a major challenge in transfer learning is to overcome the differences between the domains so that a classifier trained on the source domain generalizes well to the target domain. To overcome this problem, previous researchers proposed different methods on predominant domain generalization. For example, aiming to learn some domain-invariant features and promote good generalization across domains, the contrastive loss (Chopra et al., 2005), contrastive network (Kang et al., 2019), and adversarial training networks (Ganin Y et al., 2016; Nguyen et al., 2021) have been widely used. However, these methods were subject to the change of structure without considering the complexity that inherited within large data itself. Under this scenario, the method of continuous fine tuning was proposed to overcome this limitation. This approach was specifically designed to cover all data with only two steps, initial learning step and updated data (Kading et al., 2016). The benefits of this approach have been proved by various publications. For example, Agarwal et al. (2014) used continuously fine-tuned CNN in object recognition and achieved great improvements in performance. Zhou et al. (2021) used active, continual fine turning (ACFT) CNN to dramatically reduces annotation efforts in medical imaging by automatically selecting most representative samples. In this paper, we not only followed Kading's approach of continuous fine-tuning by adding data in more domains for improving the performance of the model on a single task but also expanding our attention to the size and sequences of data training in the experiments. Additionally, we analyze domain-relatedness and sequencing across multiple domains.

3 Data and Methods

Our approach uses three BERT-based models, RoBERTa, DistilBERT, and DeBERTa, using Amazon Reviews 2023 dataset (Hou et al., 2024) for training. We focused on three review categories: CDs & Vinyl (music), Movies & TV (films), and Grocery & Gourmet Food (food); where the first two share a similar domain while the third represents a different domain. Each model follows a different training pipeline denoted as either A, B, or C. The sequences for training and evaluation are tailored to their respective domains, ensuring targeted learning and assessment.

3.1 Datasets and Preprocessing

We leverage the Amazon Reviews 2023 dataset in this study. This is a large-scale dataset containing 571 million product reviews across 33 categories, with rich information including review text and ratings. From this corpus, we select three distinct product domains corresponding to **CDs & Vinyl**, **Movies & TV**, and **Grocery & Gourmet Food** to serve as our domains of interest. These domains (hereafter referred to as *CDs*, *Movies*, and *Food* for brevity) vary in content and vocabulary, providing a meaningful combination for both similar-domain (abbreviated Ds) and cross-domain(Dx) transfer. Each review in the dataset comes with a 5-star rating. We convert the review ratings into binary sentiment labels: reviews with 4–5 stars are labeled **positive**, and reviews with 1–2 stars are labeled **negative**, discarding 3-star "neutral" reviews to create a clear polarity distinction. We then sample a manageable subset of reviews (~1,000,~10,000, ~5,000) from each category to determine how much data is needed to achieve effective transfer learning effect. In further process, we tokenize, truncate, pad, and split data to ensures that the data is consistently formatted for input while preserving the essential content of each review for sentiment analysis.

3.2 Experimental Pipelines

Pipeline	Training Phase I	Training Phase II	Evaluation	Sample Size
a	Movies	CDs	Food	~5,000
b	CDs	Movies	Food	~1,000
c	Food	Movies	CDs	~1,000

Table 1: Data Pipelines.

To investigate the impact of dataset selection on model generalizability, we design three transfer learning pipelines A, B, and C (Table 1). Each pipeline includes two training phases followed by evaluation on a held-out domain. To test the effects of sequencing and domain similarity in a controlled manner, we: (1) reverse the fine-tuning order of two similar-domain datasets (Movies and CDs) in pipelines A(a) and B (b); and (2) use different-domain datasets (Food and Movies) for training in pipeline C (c). This comparison allows us to assess the effect of training sequencing within similar domains on transfer learning of the different domain dataset, as well as the influence of a different training sequence on datasets from different domains. Additionally, since the original sample sizes of the three datasets are comparable, we further explore whether varying the sample sizes impacts model performance by applying different scale settings across the three pipelines. The sequencing for each pipeline included 1) Training Phase I (Baseline) 2) Hyperparameter-Tuning 3) Training Phase II with secondary data and 4) Evaluation on Third Dataset.

3.3 Models and Hyperparameter Tuning

We fine-tune three different pre-trained transformer-based language models in our experiments: **roberta-base**, **distilbert-base-uncased**, and **deberta-v3-base**. These models were chosen to cover a range of model sizes and pre-training strategies, while all being state-of-the-art transformer architectures for language understanding. All three models share the same basic transformer encoder structure (multi-layer self-attention blocks) but differ in scale and pre-training approach. RoBERTa has better cross-task generalization (Liu, 2019), DeBERTa is a larger model compared to the others and has better cross domain transfer (He et al,2021) while DistilBERT has been optimized using fewer training parameters (Sanh et al, 2020).

For hyperparameter tuning, we incorporated Optuna (Akiba et al., 2019) into our training configuration for each model. We define a search space for key hyperparameters, including the learning rate, batch size, weight_decay and number of training epochs. Specifically, for each model, we conducted a modest number of trials where a model was trained with a given hyperparameter configuration, and the validation performance was recorded.

3.4 Evaluation Strategy

Three important indicators are used in our evaluation metrics, accuracy, F1-score, and validation loss. The primary evaluation indicator is classification accuracy on the positive/negative sentiment prediction on the evaluation dataset, indicating the percentage of correct predictions out of all predictions. We also compute the F1-score to balance the precision and recall. The validation loss is used to monitor model training progress and detect overfitting thus improving generalizability of the models. For each pipeline, we not only record the intermediate performance to verify that the model was indeed learning those domains and but also report the performance of the final model (after second-stage fine-tuning) on the target domain dataset. Finally, to compare performance, we use the accuracy, F1 score and validation loss, across different pipelines and models.

4 Results

We report on accuracy, F1 Score and Validation loss across the three pipelines A and three models (Appendix A, B, C).

4.1 Pipeline A

Pipeline A sequenced the data from Movies - CDs - Food. RoBERTa model baseline achieved moderate performance (accuracy: 0.827, F1-score: 0.82116) but showed inefficient learning with high validation loss (0.46662). Hyperparameter optimization using Optuna yielded modest improvement (accuracy +0.01, F1 score +0.00783, validation loss +0.14126). Noticeably, cross-domain evaluation showed RoBERTa had an increased but mild performance (accuracy: 0.83300, F1 score: 0.83041 and validation loss: 0.45873), highlighting low transfer learning capability. DistilBERT showed similar baseline performance (accuracy: 0.805, F1 score: 0.80401, validation loss: 0.60025) with insignificant gains after hyperparameter tuning. However, its cross-domain performance accuracy of 0.84680 and F1 score 0.8411 revealed greater domain sensitivity compared to RoBERTa(Figure 1).

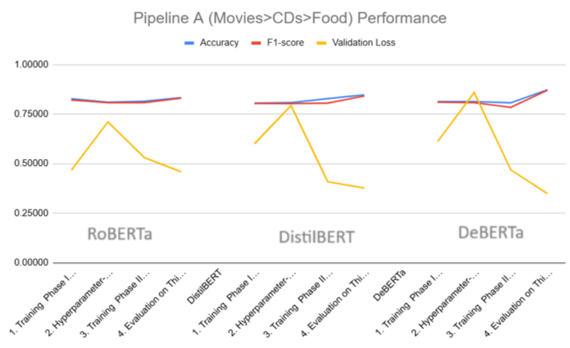

Figure 1: Evaluation metrics for Pipeline A

DeBERTa's baseline and hypertuning produced similar baseline performance (accuracy: 0.813) but higher validation loss (0.8594) after hypertuning indicating potential overfitting. While secondary domain training decreased the accuracy to 0.807, the validation loss was halved to 0.46819 suggesting improved generalization. DeBERTa significantly performed well in cross-domain evaluation (accuracy: 0.8724, F1 score:0.87015 and, validation loss: 0.34902), overshadowing RoBERTa's and DistilBERT's transfer learning performance. This pattern indicates that while DeBERTa may require careful tuning to avoid overfitting on source domains, it develops robust representations that transfer effectively to new domains. All three models demonstrated the importance of validation loss as a key indicator of generalization potential beyond standard accuracy metrics.

4.2 Pipeline B

Pipeline B trained on same datasets as pipeline A with a different order and evaluated on the dissimilar dataset (CDs-Movies-Food). RoBERTa demonstrated exceptional in-domain performance on music reviews (accuracy: 0.970, F1: 0.971, loss: 0.079), with further optimization pushing metrics near perfection (accuracy: 0.972, F1: 0.97131, and loss: 0.10217). The model maintained similar high performance when transferred to the related movies domain with the validation loss dropping to 0.078. However, interestingly, showed degradation on dissimilar grocery data with accuracy at 0.925, F1 score at 0.92206 and loss of 0.212(Figure 2).

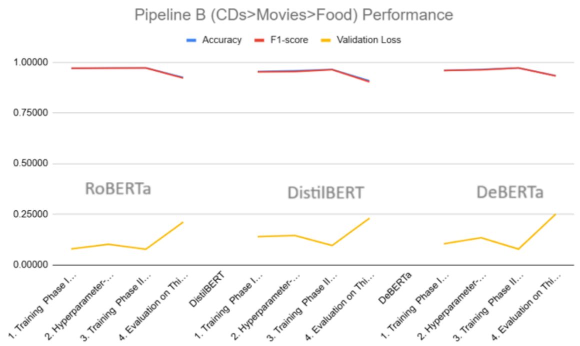

Figure 2: Evaluation metrics for Pipeline B

DistilBERT and DeBERTa followed similar trajectories in initial training with DeBERTa leading with 0.95982 accuracy and F1 score 0.95994. Both maintain strong performance during entertainment-domain transfer with DeBERTa achieving 0.972 accuracy and DistillBERT at 0.96422. However, on evaluation using the grocery review, performance dropped across all the models with accuracy on RoBERTa, DistilBERT and DeBERTa having 0.92493, 0.90824, and 0.93385 respectively.

4.3 Pipeline C

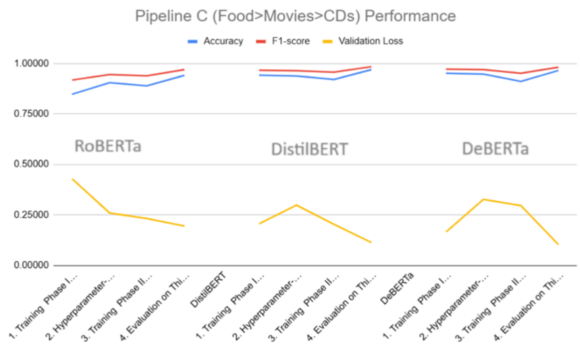

Figure 3: Evaluation metrics for Pipeline C

Pipeline C had a reverse domain progression (Grocery-Movies-Music) to better show variation of sequencing for transfer learning. RoBERTa demonstrated strong initial performance on food reviews (accuracy: 84.7%, F1: 0.917), with hyperparameter tuning significantly boosting results (accuracy: 0.905, F1: 0.945). However, cross-domain fine-tuning on movie data caused performance degradation to an accuracy of 0.889), demonstrating challenges in domain-shifting.The final evaluation on music reviews, a domain similar to the last training dataset, movies but nonetheless unseen during training yielded very high performance (accuracy: 0.941, F1: 0.969, loss: 0.19477)(Figure 3). This suggests RoBERTa effectively performs well with similar domain knowledge as opposed to dissimilar domains.

DistilBERT and DeBERTa outperformed RoBERTa in initial food reviews. DeBERTa led with 0.951 accuracy, but showed similar cross-domain struggles during training with movies dataset with a decrease of accuracy and F1 scores to 0.91115 and 0.95109 respectively, and an increase of validation loss to 0.29576. However, on evaluation using CDs data, there was a sharp increase in the metrics (accuracy: 0.96484, F1 score: 0.98118 and loss: 0.10297) showing incredible ability to drastically improve similar domain learning. DistilBERT also demonstrated ability for transfer learning with an evaluation accuracy of 0.96923, F1 score of 0.98361 and low validation loss of 0.11316 compared to the training metrics (accuracy: 0.92023, F1 score:0.95646,loss: 0.20338) thus demonstrating high performance for transfer learning as well.

5 Discussion

To allow for comparisons across the pipelines, the percentage of increase or decrease of accuracy across the models was computed from Training Phase II. This enabled comparative analysis to determine which pipelines demonstrated transfer learning, and the best performing model in each pipeline. We additionally contrast the performance of the models based on dataset sizing.

Model performance: Our results reveal DeBERTa's superior cross-domain adaptability, demonstrating the highest accuracy improvements (+6.54% for Pipeline A's movie/TV domain and +5.37% for Pipeline C's dissimilar grocery/food domain) while maintaining the smallest accuracy degradation (-3.84% in Pipeline C) (Table 2). These findings confirm He et al's (2021) research on DeBERTa's advanced architecture and effectiveness in cross-domain as cross domain transfer. This evidence positions DeBERTa as the most optimal choice for cross-domain transfer tasks, especially for low resource datasets, however its computational demands are high compared to other BERT models.

Cross-Model Generalization: The pipelines highlighted DeBERTa's relative advantage in cross-domain tasks. All models exhibited decreasing validation loss patterns confirming learning and high generalizability across all models and pipelines. Additionally, in pipelines A and C, all models exhibited remarkably when evaluated for transfer learning with evaluation accuracy increasing across the 3 models. This pattern confirms four critical insights: (1) domain similarity significantly impacts transfer success more than absolute model performance, and (2) initial strong performance doesn't guarantee cross-domain stability, 3)Models trained on multiple types of datasets could transfer learning to a new, never seen dataset exceptionally well when the new dataset was in a similar domain to any of the training datasets, and 4) Exposure to more data, even from different domains, appeared to strengthen the model's overall prediction and generalization capability. The results highlight DeBERTa's advantage in maintaining consistent performance across domain shifts.

Hyperparameter tuning and dataset sizing: There was significant variation in how models respond to hyperparameter tuning across the three different pipelines (Table 3). RoBERTa shows the most consistent improvements, with accuracy gains in all pipelines, recording an increase of 5.71% for Pipeline C, which had the smallest dataset. This suggests that RoBERTa responds well to fine tuning with low resource datasets. While both improve marginally in Pipelines A and B (~0.3–0.44%), DistilBERT and DeBERTa suffer slight declines in Pipeline C (–0.38% and –0.45%, respectively), implying inconsistency. As to the data size effect, across

Pipeline	RoBERTa	DistilBERT	DeBERTa
A(~5,000)	1.8%	1.88%	6.54%
B(~10,000)	-4.74%	-5.60%	-3.84%
C(~1,000)	5.16%	4.9%	5.37%

Table 2: Change of accuracy for tertiary dataset across three models and pipelines.

Pipeline	RoBERTa	DistilBERT	DeBERTa
A(~5,000)	1.00%	0.30%	0.00%
B(~10,000)	0.22%	0.33%	0.44%
C(~1,000)	5.71%	-0.38%	-0.45%

Table 3: Change of accuracy after Hyperparameter tunning across three models and pipelines.

larger datasets in Pipelines A and B with ~5000 and ~10000 rows respectively, all models show modest gains, reinforcing that tuning's value scales with data volume with the possibility of diminishing returns in large datasets. These trends suggest model selection for tuning should consider dataset size, with RoBERTa preferred for sparse data and conversely DistilBERT and DeBERTa models requiring more focus when fine tuning. Additionally, pipeline A (5000) and pipeline C(1000) both show transfer learning leading to the conclusion that smaller datasets are adequate for transfer learning, which is normally the case with low-resource datasets. These findings align with Yosinski et al. (2014) and Vu et al. (2020), who stress the relevance of feature stability and task similarity. However, we extend prior work by showing that even minimal data (1,000 examples) can yield strong results when the sequence and model are carefully chosen.

Dataset sequencing: The experiments across Pipelines A, B, and C have shown the impacts of dataset sequencing to transfer learning performance. When models were trained on related domains (movies and music) before being evaluated on dissimilar data (food reviews), we observed performance improvement in Pipeline A indicating cross domain transfer learning and degradation while Pipeline B lack of transfer learning. The variation was due to differences in sequencing of the similar datasets in both pipelines. This suggests transfer learning works better when moving from broad (movies) to specific domains (music and food) rather than vice versa. However, for Pipeline A and C, DeBERTa showed slightly better cross-domain retention, maintaining 2-3% higher accuracy than RoBERTa and DistilBERT in these tests. These results demonstrate that while training on semantically related domains helps improve cross domain transfer learning performance the modest increases in performance suggest domain differences still pose challenges regardless of model architecture.

The order of domain exposure proved particularly important, as shown by Pipeline C's reverse sequence (food - movies - music). This shows that data should be sequenced strategically, prioritizing related domains later in training to maximize generalization. The performance of the pipelines demonstrate that successful transfer learning depends more on thoughtful domain sequencing and model selection than raw performance metrics, with validation loss serving as a crucial indicator of true generalization capability.

6 Limitation

This study is not free from limitations. First, although the datasets represent different product categories, all data sources consist of consumer-generated reviews. As such, the linguistic differences across domains may not be substantial enough to fully evaluate the challenges of domain adaptation. Second, due to computational constraints, the models were trained on two datasets only. This limitation may have hindered the models' ability to learn broadly generalized features, particularly in cross-domain settings. Finally, because the datasets are derived from real-world reviews, they may contain class imbalances and other inherent biases that can skew model performance. Addressing these biases through data preprocessing or algorithmic approaches remains an important direction for future research.

7 Conclusion

Our experiments show that transfer learning success depends more on domain relationships than absolute model performance, with data sequencing acting as a key contributor in maximizing results. The study provides a systematic comparison of BERT-family models in transfer learning for NLP classification. We find that DeBERTa is the most robust model for domain adaptation, dataset relatedness and sequencing significantly influence performance, a sample training size of less than 1,000 is often sufficient to yield measurable transfer benefits. Our pipeline-based approach offers a template for future research into sequence-aware transfer learning strategies, which could expand this evaluation to multilingual settings and additional NLP tasks. Additionally, expanded testing on hybrid sequences such as alternating domains could provide further insights into optimize transferability.

References

Agarwal S, Terrail JO, Jurie F. Recent advances in object detection in the age of deep convolutional neural networks. arXiv preprint arXiv:1809.03193. 2018 Sep 10.

Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A next-generation hyperparameter optimization framework. InProceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining 2019 Jul 25 (pp. 2623-2631).

Chopra, S., Hadsell, R., & LeCun, Y. (2005, June). Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)* (Vol. 1, pp. 539-546). IEEE.

Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. InProceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers) 2019 Jun (pp. 4171-4186).

Ganin, Y., et al. (2016). Domain-Adversarial Training of Neural Networks. JMLR.https://jmlr.org/papers/volume17/15-239/15-239.pdf

Gholizade M, Soltanizadeh H, Rahmanimanesh M, Sana SS. A review of recent advances and strategies in transfer learning. International Journal of System Assurance Engineering and Management. 2025 Feb 21:1-40.

He, P., et al (2021). "*DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing"* ArXiv.org , 24 Mar 2023, https://arxiv.org/pdf/2111.09543

Hou Y, Li J, He Z, Yan A, Chen X, McAuley J. Bridging language and items for retrieval and recommendation. arXiv preprint arXiv:2403.03952. 2024 Mar 6.

Howard, J., & Ruder, S. (2018). "Universal Language Model Fine-tuning for Text Classification (ULMFiT)."https://arxiv.org/abs/1801.06146

Käding C, Rodner E, Freytag A, Denzler J. Fine-tuning deep neural networks in continuous learning scenarios. InAsian Conference on Computer Vision 2016 Nov 20 (pp. 588-605). Cham: Springer International Publishing.

Kang G, Jiang L, Yang Y, Hauptmann AG. Contrastive adaptation network for unsupervised domain adaptation. InProceedings of the IEEE/CVF conference on computer vision and pattern recognition 2019 (pp. 4893-4902).

Kouw, Wouter M, and Marco Loog, "An introduction to domain adaptation and transfer learning", ArXiv.org, 31 Dec 2018,, https://arxiv.org/abs/1812.11806

Li, X., et al. (2021). Measuring Transferability in Cross-Domain Learning. NeurIPS.https://proceedings.neurips.cc/paper_files/paper/2022/file/11b3ae28275461741026c46c0c786711-Paper-Conference.pdf

Liu, T.,et a (2019). "*RoBERTa: A Robustly Optimized BERT Pretraining Approach*", ArXiv.org , 26 Jul 2019, https://arxiv.org/pdf/1907.11692

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). *Learning word vectors for sentiment analysis*. arXiv preprint arXiv:1103.0398. IMDB Dataset

Misra, R., & Arora, P. (2019). *News Headlines Dataset for Sarcasm Detection*. GitHub Repository

Nguyen, A. T., Tran, T., Gal, Y., & Baydin, A. G. (2021). Domain invariant representation learning with domain density transformations. *Advances in Neural Information Processing Systems*, *34*, 5264-5275.

Sanh, V.,et al (2020),"*DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*", ArXiv.org, 1 Mar 2020, https://arxiv.org/pdf/1910.01108

Tzeng, E., et al. (2017). Adversarial Discriminative Domain Adaptation. CVPR.https://openaccess.thecvf.com/content_cvpr_2017/papers/Tzeng_Adversarial_Discriminative_Domain_CVPR_2017_paper.pdf

Vu, Tu, et al. "Exploring and Predicting Transferability across NLP Tasks." ArXiv.org, 6 Oct. 2020, arxiv.org/abs/2005.00770

Xiang Zhang, and Acharki Yassir. (2022). Amazon Reviews for SA fine-grained 5 classes [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/3499094

Xiao,F. et al, "Transductive Learning for Unsupervised Text Style Transfer", ArXiv.org, 16 Sep 2021, https://arxiv.org/abs/2109.07812

Yada Pruksachatkun et al." Intermediate-Task Transfer Learning with Pretrained Language Models: When and Why Does It Work?" aclanthology.org, 2020, aclanthology.org/2020.acl-main.467/

Yesmin,J.(2024)."*Cross-Domain Evaluation for Multi-Task Learning in NLP: A Unified Framework for Generalization and Robustness*",https://papers.ssrn.com/ , 2025, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5018566

Yosinski,J. Et al. "How transferable are features in deep neural networks?", ArXiv.org, 6 Nov. 2024, https://arxiv.org/pdf/1411.1792

You, K., et al. (2021). LogME: Practical Assessment of Model Transferability. ICML.https://arxiv.org/pdf/2102.11005

Zhao, Z., Alzubaidi, L., Zhang, J., Duan, Y., & Gu, Y. (2024). A comparison review of transfer learning and self-supervised learning: Definitions, applications, advantages and limitations. *Expert Systems with Applications*, *242*, 122807.

Zhou Z, Shin JY, Gurudu SR, Gotway MB, Liang J. Active, continual fine tuning of convolutional neural networks for reducing annotation efforts. Medical image analysis. 2021 Jul 1;71:101997.

Author Contribution

All authors were involved in the conceptualization of the project. All authors developed the experimental design, initiated data analysis, built and fine-tuned three models. Hildah performed the analysis with pipeline A, Peng with pipeline B, and Qiong with pipeline C. All authors prepared the draft manuscript and reviewed and edited the manuscript. All authors read and approved the final manuscript. All authors prepared the slides.

Appendices

Appendix A. RoBERTa Results.

RoBERTa	Pipeline	Dataset	Accuracy	F1-score	Validation Loss
1. Training Phase I (Baseline)	A	Movies and TV	0.82700	0.82116	0.46662
1. Training Phase I (Baseline)	B	CD and Vynl	0.97028	0.97073	0.07944
1. Training Phase I (Baseline)	C	Grocery and Gourmet Food	0.84748	0.91745	0.42861
2. Hyperparameter-Tuning	A	Movies and TV	0.83700	0.82899	0.60788
2. Hyperparameter-Tuning	B	CD and Vynl	0.97248	0.97131	0.10217
2. Hyperparameter-Tuning	C	Grocery and Gourmet Food	0.90458	0.94466	0.25960
3. Training Phase II with secondary data	A	CD and Vynl	0.81500	0.80744	0.52955
3. Training Phase II with secondary data	B	Movies and TV	0.97228	0.97197	0.07788
3. Training Phase II with secondary data	C	Movies and TV	0.88911	0.93864	0.23236
4. Evaluation on Third Dataset	A	Grocery and Gourmet Food	0.83300	0.83041	0.45875
4. Evaluation on Third Dataset	B	Grocery and Gourmet Food	0.92493	0.92206	0.21175
4. Evaluation on Third Dataset	C	CD and Vynl	0.94066	0.96871	0.19477

Appendix B. DistilBERT Results.

DistilBERT	Pipeline	Dataset	Accuracy	F1-score	Validation Loss
1. Training Phase I (Baseline)	A	Movies and TV	0.80500	0.80401	0.60025
1. Training Phase I (Baseline)	B	CD and Vynl	0.95432	0.95227	0.14023
1. Training Phase I (Baseline)	C	Grocery and Gourmet Food	0.94140	0.96573	0.20638
2. Hyperparameter-Tuning	A	Movies and TV	0.80800	0.80263	0.79287
2. Hyperparameter-Tuning	B	CD and Vynl	0.95762	0.95375	0.14507
2. Hyperparameter-Tuning	C	Grocery and Gourmet Food	0.93764	0.96380	0.29860
3. Training Phase II with secondary data	A	CD and Vynl	0.82800	0.80530	0.40868
3. Training Phase II with secondary data	B	Movies and TV	0.96422	0.96363	0.09590
3. Training Phase II with secondary data	C	Movies and TV	0.92023	0.95646	0.20338
4. Evaluation on Third Dataset	A	Grocery and Gourmet Food	0.84680	0.84110	0.37663
4. Evaluation on Third Dataset	B	Grocery and Gourmet Food	0.90824	0.90367	0.23066
4. Evaluation on Third Dataset	C	CD and Vynl	0.96923	0.98361	0.11316

Appendix B. Pipeline B Results.

DeBERTa	Pipeline	Dataset	Accuracy	F1-score	Validation Loss
1. Training Phase I (Baseline)	A	Movies and TV	0.81300	0.80978	0.61047
1. Training Phase I (Baseline)	B	CD and Vynl	0.95982	0.95994	0.10457
1. Training Phase I (Baseline)	C	Grocery and Gourmet Food	0.95116	0.97140	0.16622
2. Hyperparameter-Tuning	A	Movies and TV	0.81300	0.80786	0.85894
2. Hyperparameter-Tuning	B	CD and Vynl	0.96423	0.96311	0.13405
2. Hyperparameter-Tuning	C	Grocery and Gourmet Food	0.94666	0.96898	0.32674
3. Training Phase II with secondary data	A	CD and Vynl	0.80700	0.78429	0.46819
3. Training Phase II with secondary data	B	Movies and TV	0.97228	0.97197	0.07838
3. Training Phase II with secondary data	C	Movies and TV	0.91115	0.95109	0.29576
4. Evaluation on Third Dataset	A	Grocery and Gourmet Food	0.87240	0.87015	0.34902
4. Evaluation on Third Dataset	B	Grocery and Gourmet Food	0.93385	0.93281	0.25127
4. Evaluation on Third Dataset	C	CD and Vynl	0.96484	0.98118	0.10297