

MULTI-SOURCE DOMAIN GENERALIZATION FOR TRANSFORMER-BASED SENTIMENT CLASSIFICATION WITH LIGHTWEIGHT REPRESENTATION REFINEMENT

A THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE AWARD OF THE DEGREE
OF

MASTER OF TECHNOLOGY

IN
ARTIFICIAL INTELLIGENCE

Submitted by

ROHIT KUMAR (2K24/AFI/30)

Under the supervision of

Prof. Aruna Bhat



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY**

(Formerly Delhi College of Engineering)

Bawana Road, Delhi 110042

MAY, 2026

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi-110042

CANDIDATE'S DECLARATION

I, **Rohit(2K24/AFI/30)**, hereby certify that the work which is being presented in the major project report II entitled **“Multi-Source Domain Generalization for Transformer-Based Sentiment Classification with Lightweight Representation Refinement”** in partial fulfillment of the requirements for the award of the Degree of Master of Technology, submitted in the **Department of Computer Science and Engineering**, Delhi Technological University is an authentic record of my own work carried out during the period from Start Date, e.g., August 2024 to May 2026 under the supervision of **Prof. Aruna Bhat**.

The matter presented in this thesis has not been submitted by me for the award of any other degree of this or any other Institute.

Rohit Kumar

Candidate's Signature

This is to certify that the student has incorporated all the corrections suggested by the examiners in the thesis and the statement made by the candidate is correct to the best of our knowledge.

Aruna Bhat

Signature of Supervisor

Signature of External Examiner

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

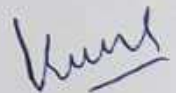
Bawana Road, Delhi-110042

CERTIFICATE

I hereby certify that the Project titled **“Multi-Source Domain Generalization for Transformer-Based Sentiment Classification with Lightweight Representation Refinement”**, submitted by **Rohit Kumar**, Roll No. **2K24/AFI/30**, Department of **Computer Science & Engineering**, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of the degree of **Master of Technology (M.Tech)** in Artificial Intelligence is a genuine record of the project work carried out by the student under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree to this University or elsewhere.

Place : Delhi

Date : 24.05.2026



Dr. Aruna Bhat

Professor

Delhi Technological University

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi-110042

ACKNOWLEDGEMENT

I would like to express my sincere gratitude and appreciation to my supervisor, **Prof. Aruna Bhat**, Department of Computer Science and Engineering, Delhi Technological University, for her continuous guidance, valuable insights, and constant encouragement throughout the course of this project. Her expertise in the field and her willingness to support me at every stage of the research were instrumental in shaping this work.

I am also deeply thankful to the faculty members of the Department of Computer Science and Engineering at Delhi Technological University for their academic support and for fostering an environment that encourages research and inquiry. Their courses and discussions have helped me build the foundation on which this thesis is based.

Place: Delhi
Date: 24.05.2026

ROHIT KUMAR
(24/AFI/30)

ABSTRACT

Fine-tuned encoders based on the transformer architecture produce extremely accurate results when testing review texts are drawn from the same product category as training texts. Nevertheless, the models that are applied to review texts from another product category show lower accuracy, inconsistency in predictions between the initial states of the model, and unpredictable behavior. Domain adaptation addresses this issue by adding samples belonging to the target product category into the training process; however, a practical sentiment analysis system that can be deployed in production lacks access to target category samples. There might simply be no category to review yet at the moment when the model is being developed. The more complicated problem, namely, training on any labelled categories and later producing good results on an unseen one, is called domain generalization. This is the topic of my thesis.

Below is the empirical study performed on multi-source domain generalization of sentiment classification task using transformer-based neural networks on Amazon Multi-Domain Sentiment dataset. In all experiments, the strict rule of no-target access will be maintained: no labels and even unlabelled texts from the target domain will be used during the training phase. The DistilBERT base uncased architecture was chosen as the encoder in my experiment, and I compare four setups: (i) empirical risk minimization (ERM) using single source – training of the model separately for each category; (ii) ERM using multiple sources – merging three categories into one dataset; (iii) DANN baseline for multiple sources; and (iv) the approach suggested in my thesis with adding a tiny projection-residual module on top of the DistilBERT CLS vector while keeping the encoder fully fine-tuned. This additional module is based on the two-layer bottleneck architecture with a skip connection. It takes roughly 0.6% of total model parameters, and initially serves as an identity function.

With leave-one-domain-out on the four domains (Books, DVD, Kitchen, and Electronics), our proposed model reaches an average cross-domain accuracy of 86.56%. This number exceeds the single-source model baseline by 1.40%, beats the multi-source ERM baseline by 0.69%, and surpasses the DANN baseline by 0.50%. The most notable gains are made in target domains Books and Kitchen, where the proposed approach outperforms the multi-source ERM baseline in all random seed configurations. In the challenging target domain Electronics, the projection-residual module does not contribute much to the average, but almost halves the standard deviation (0.54 to 0.31) through a more stable performance

pattern—a finding verified through a second residual-connection ablation analysis. With another ablation test, we found out that the projection-residual module is not particularly sensitive to the bottleneck dimension size r . We experimented with bottleneck dimensions $r = \{64, 128, 256, 512\}$, and observed that all four options yielded similar average accuracies within one percentage point of each other, except for $r = 64$, which performed better in the Electronics target domain.

This research concludes that a simple combination of all-source category aggregation, using cross-entropy fine-tuning, and using an additional refinement layer on top of the encoder's classification token is sufficient for obtaining significant and reproducible performance gains in cross-domain sentiment classification. It is not necessary to use any adversarial training or invariance regularization, nor any modifications to the encoder's structure.

Keywords: domain generalization, cross-domain sentiment classification, multi-source training, transformer models, DistilBERT, representation refinement, residual learning, robustness, Amazon reviews.

CONTENTS

Candidate's Declaration	ii
Certificate	iii
Acknowledgement	iv
Abstract.....	v
Contents	vii
List of Tables	ix
List of Figures.....	x
List of Abbreviations	xi
CHAPTER 1 INTRODUCTION.....	1
1.1 Background and Motivation.....	1
1.2 Problem Statement.....	3
1.3 Research Objectives.....	3
1.4 Research Questions.....	4
1.5 Contributions.....	4
1.6 Scope and Limitations.....	5
1.7 Organization of the Thesis	6
CHAPTER 2 LITERATURE REVIEW.....	7
2.1 Distribution Shift in Machine Learning.....	7
2.2 Domain Adaptation.....	8
2.3 Domain Generalization	9
2.4 Multi-Source Learning.....	10
2.5 Transformer Models for Sentiment Classification.....	10
2.6 Regularization and Robustness in Transformers.....	11
2.7 Lightweight Architectural Refinement	12
2.8 Causal and Representation-Based Approaches.....	13
2.9 Research Gap	13
CHAPTER 3 PROPOSED METHODOLOGY	15
3.1 Problem Formulation	15
3.2 Backbone Encoder: DistilBERT	16
3.3 Multi-Source Training Strategy	17
3.4 Lightweight Projection Residual Module	17
3.5 Sentiment Classification Head and Objective.....	19

3.6 Overall Architecture and Algorithm	20
3.7 Implementation Details.....	21
CHAPTER 4 EXPERIMENTAL SETUP	23
4.1 Dataset.....	23
4.2 Preprocessing Pipeline	24
4.3 Evaluation Protocol: Leave-One-Domain-Out	24
4.4 Baselines	25
4.5 Metrics	26
4.6 Statistical Validation	27
CHAPTER 5 RESULTS AND ANALYSIS	28
5.1 Single-Source Baseline	28
5.2 Multi-Source ERM.....	29
5.3 DANN Baseline	30
5.4 Proposed Method: Multi-Source + Projection Residual	31
5.5 Full Cross-Method Comparison.....	32
5.6 Variance and Stability Analysis.....	33
5.7 Ablation Study — Bottleneck Dimension.....	35
5.8 Ablation Study — Residual Connection	36
5.9 Qualitative Discussion	37
5.10 Synthesis	38
CHAPTER 6 CONCLUSION AND FUTURE WORK.....	40
6.1 Summary of Findings.....	40
6.2 Contributions Revisited.....	41
6.3 Limitations	42
6.4 Future Work	42
6.5 Closing Remarks.....	43
References	44

LIST OF TABLES

Table 4.1	Dataset statistics for the four domains.....	23
Table 5.1	Single-Source ERM accuracy and F1.....	28
Table 5.2	Multi-Source ERM accuracy and F1.....	29
Table 5.3	DANN accuracy and F1	30
Table 5.4	Proposed method accuracy and F1	31
Table 5.5	Leave-one-domain-out accuracy (all methods)	32
Table 5.6	Leave-one-domain-out F1 (all methods)	32
Table 5.7	Per-seed accuracy of MS-ERM vs Proposed.....	34
Table 5.8	Ablation of the bottleneck dimension	35
Table 5.9	Ablation of the residual connection.....	36

LIST OF FIGURES

Figure 3.1 Architecture of the proposed multi-source DG framework	20
Figure 5.1 Leave-one-domain-out accuracy by method.....	33
Figure 5.2 Cumulative improvement chain	33
Figure 5.3 Per-seed accuracy stability.....	34
Figure 5.4 Bottleneck-dimension ablation curve.....	35
Figure 5.5 Residual-connection ablation	36
Figure 5.6 Training loss curves	37
Figure 5.7 PCA of [CLS] / refined representations	38
Figure 5.8 Confusion matrices for proposed method	38

LIST OF ABBREVIATIONS

Abbreviation	Full Form
NLP	Natural Language Processing
DG	Domain Generalization
DA	Domain Adaptation
ERM	Empirical Risk Minimization
DANN	Domain-Adversarial Neural Network
IRM	Invariant Risk Minimization
DRO	Distributionally Robust Optimization
LODO	Leave-One-Domain-Out
MS-ERM	Multi-Source Empirical Risk Minimization
BERT	Bidirectional Encoder Representations from Transformers
CLS	Classification (token in BERT)
MLP	Multi-Layer Perceptron
ReLU	Rectified Linear Unit
GRL	Gradient Reversal Layer
GPU	Graphics Processing Unit
F1	F-measure (harmonic mean of precision and recall)
PCA	Principal Component Analysis

CHAPTER 1

INTRODUCTION

1.1 Background and Motivation

Sentiment classification has been one of the staple tools of contemporary natural language processing. Under the hood of consumer reviews aggregation on all large-scale e-commerce websites, of perception monitoring dashboards within marketing departments, and behind the curtains of various trend detection algorithms that analyze social media discourse about a brand lies the task of identifying the polarity of each message. Over the last several years, a particular class of models has come to dominate this space, namely transformer encoders, be it BERT [9], RoBERTa [21], DeBERTa [22], or even their compact alternatives like DistilBERT [10], combined with a lightweight classifier head trained using labeled data.

It is clear how attractive such an approach would be for any practitioner. Just a few hundreds of thousands of labeled examples, an already trained encoder, three to four passes through the data, and the generated classifier will inevitably reach above 90% accuracy on test data drawn from the same distribution. However, things become challenging once the distribution where we intend to deploy our model differs from that on which it was trained. The model of a book review may fail to work on a review of a headphone not due to any complex language constructions and subtle nuances, but simply because the context itself has changed. The word “gripping” works well for a thriller and poorly for a bicycle handlebar. “Light” is associated with good luggage and bad coffee. Consumers of electronic devices tend to be laconic and analytical whereas literary consumers are likely to be verbose and expressive. All of this is certainly not a speculation. Ben-David et al. [1] formalized the problem more than a decade ago bounding the error made by a classifier on a target distribution by (i) its error on the source distribution used for training and (ii) the divergence between those distributions.

Adaptation is the natural response to such changes. A wide variety of approaches fall into this category – adversarial alignment [2], correlation alignment [3], and many more – and all of them operate under the assumption that a sample of the target data, possibly unlabeled, can be obtained during the learning process. While this may hold for many

academic datasets, it holds far less true for many real-world cases. A group designing a sentiment analysis system for a new product category usually has neither the labeled examples of the old dataset nor an unlabeled set to extract features from; they have a time frame that makes months of collecting new data impractical, and their only option is the labeled data for old product categories. This case, when the only data is the labeled one that the team already has, while the target is unseen, is now known as the problem of domain generalization [7].

There have been many advances in the field of domain generalization over the past five years. Some of these papers introduce new loss functions (Invariant Risk Minimization [4], Group Distributionally Robust Optimization [11]), others introduce new data augmentation techniques (MixStyle [14] and other variants), while yet others propose entirely new architectures. However, an emerging and rather disillusioning result from this body of research indicates that none of these sophisticated frameworks has succeeded in beating a well-configured baseline algorithm. For example, the DomainBed study conducted by Gulrajani and Lopez-Paz [5] evaluated a broad cross-section of the domain-generalization toolkit using a standardized evaluation protocol and found that Empirical Risk Minimization over all the domains, when pooled together, performs competitively – and often beats those techniques that were explicitly devised to beat it.

That consensus is taken seriously in the way the method described in Chapter 3 has been designed. Instead of designing new loss functions or invariance penalties, I rely on standard multi-source empirical risk minimisation, which has been demonstrated recently to perform exceptionally well and is considered in this context as the strong baseline to beat. The originality of the proposed solution lies in the architecture of the model used for learning. Rather than feeding the representation produced by DistilBERT following the encoding of a review into a linear classifier, it is first passed through a lightweight refinement network consisting of two layers with a residual skip connection. That refinement network is trained during multi-source fine-tuning to learn how to guide representations towards characteristics that are likely to generalise across product classes. The design of this refinement network is intentionally minimal: in terms of parameters, it only adds about 0.6% of parameters to those of DistilBERT. It has been inspired by the notion of adapter modules [15] but is designed to serve a different purpose.

The questions I want to answer with this setup are concrete. How much does the cross-domain gap actually cost, in percentage points of accuracy, when a transformer is trained on one product category and tested on another? Does pooling several source categories close that gap in a way that holds up across random seeds? Does the lightweight refinement module add anything beyond what pooling alone provides? And does the proposed approach beat an adversarial alignment baseline — DANN, the standard adversarial DG method — when both are run under identical training conditions? Chapter 5 answers each of these questions with mean-and-standard-deviation numbers from a leave-one-domain-out evaluation over three random seeds per configuration.

1.2 Problem Statement

Stripped down to its essentials, the problem this thesis addresses is the following. A practitioner is given labelled review data from several product categories — say, books, DVDs, and kitchen items — and is asked to ship a sentiment classifier that will be deployed on reviews from a fourth, completely different category, say electronics. No labelled review from the electronics category is available. No unlabelled crawl of electronics reviews has been collected either. The training pipeline must commit to a model based purely on the three source categories, and the model so produced must generalise. The task is to design a training procedure and a small architectural refinement that, taken together, raise the expected accuracy and reduce the seed-to-seed variability of the classifier on that unseen target category, relative to the naive baseline of fine-tuning a transformer on whichever source happens to be available.

1.3 Research Objectives

Four concrete objectives organise the rest of the thesis.

1. To measure how much cross-domain accuracy a DistilBERT-based sentiment classifier loses when it is fine-tuned on a single source category and then evaluated on a different category from the Amazon Multi-Domain Sentiment Dataset. The measurement is to be reported as mean and standard deviation over multiple random seeds, so that the size of the cross-domain gap can be distinguished from seed-induced noise.
2. To establish whether pooling three source categories during fine-tuning — without any change to the architecture, the loss, or the optimisation procedure — improves

cross-domain accuracy on a fourth, unseen target category, and to characterise the per-target pattern of that improvement.

3. To design a lightweight projection-residual module that sits on top of DistilBERT's [CLS] vector, to specify its forward pass and its parameter budget, and to evaluate empirically whether adding this module to a multi-source pipeline produces a further improvement over multi-source ERM alone.
4. To validate the proposed approach by comparing it against the Domain-Adversarial Neural Network baseline of Ganin and Lempitsky [2] under matched training data and matched compute, in order to determine whether the simpler architectural refinement is competitive with — or superior to — the standard adversarial-alignment alternative.

1.4 Research Questions

The objectives above can be restated as four research questions that Chapter 5 will return to and answer with empirical numbers.

RQ1. Given a DistilBERT-based classifier for sentiment analysis, which is trained with one domain from the Amazon Multi-Domain Sentiment Dataset and tested against an unseen domain, how much of a loss in cross-domain accuracy can be observed, and how much variance exists in the loss across domains?

RQ2. Does pooling three source categories during fine-tuning, without altering the loss or the architecture, produce a measurable and seed-reliable improvement in cross-domain accuracy over the corresponding single-source baseline?

RQ3. Does the lightweight projection-residual module proposed in Chapter 3, applied on top of the [CLS] vector of a fully fine-tuned DistilBERT trained with multi-source ERM, deliver an additional improvement over the multi-source ERM baseline, and does it tighten the spread of accuracies across random seeds?

RQ4. When the proposed multi-source-plus-refinement pipeline is compared against the DANN adversarial-alignment baseline under matched conditions, which approach performs better on the leave-one-domain-out protocol — on the mean and on the seed-to-seed variance?

1.5 Contributions

The work presented in this thesis makes five contributions.

1. *An end-to-end empirical study of multi-source domain generalization for transformer-based sentiment classification under strict no-target-access conditions.* The study uses a leave-one-domain-out protocol over four product categories of the Amazon Multi-Domain Sentiment Dataset, with three random seeds per configuration, and reports per-target accuracy and F1 with mean and standard deviation for all four methods compared.
2. *Quantitative support for unevenness of effectiveness of source pooling.* Source pooling of three source types increases the average accuracy on leave-one-domain-out classification by 0.72 percent points compared to the single-source baseline (from 85.15% to 85.87%). However, this entire increase is only attributable to the most difficult transition from source to target domains (target=Electronics, increase due to pooling = 3.17 percentage points).
3. *A lightweight projection-residual module for [CLS] refinement.* The module is a two-layer bottleneck with a ReLU non-linearity and a residual skip connection, adding roughly 394,000 trainable parameters (approximately 0.6% of DistilBERT's 66M) at the default bottleneck dimension $r = 256$. Added on top of the multi-source pipeline, the module raises average cross-domain accuracy a further 0.69 percentage points to 86.56%, with the largest improvements on Books (+1.13 over MS-ERM, robust across all three seeds) and Kitchen (+1.35, robust across all three seeds).
4. *A controlled head-to-head against DANN.* Run on the same multi-source pool, with the same encoder, batch size, learning rate, and number of epochs, DANN improves over MS-ERM by only 0.19 points on average — within seed-induced variance — and the simpler proposed approach beats DANN by 0.50 points on average.
5. *Two ablations that isolate the contributions of the module's design choices.* A sweep of the bottleneck dimension $r \in \{64, 128, 256, 512\}$ on the Electronics target shows that the module is insensitive to r within roughly one accuracy point, with $r = 64$ (only ~99,000 added parameters) the narrowly best setting. A direct comparison of the module with and without its residual skip shows that the skip leaves mean accuracy essentially unchanged but reduces seed-to-seed variance by roughly a factor of two — confirming that the residual functions as a stability mechanism rather than as an accuracy mechanism.

1.6 Scope and Limitations

The work is deliberately narrow in several respects, which is worth stating explicitly so that the claims of the thesis are not over-read.

- **Task.** Only binary sentiment (positive versus negative) is considered. Fine-grained ratings, aspect-based sentiment, and emotion classification are out of scope. The methodology is dataset-agnostic and could be transported to those settings, but no experimental claim about them is made here.
- **Language.** All text in the dataset is English. Whether the source-pooling-plus-refinement recipe transfers to multilingual sentiment is not tested.
- **Dataset.** Experiments use the Amazon Multi-Domain Sentiment Dataset [8] and only that dataset. The four product categories used (Books, DVD, Kitchen, Electronics) share the genre of consumer review but differ substantially in vocabulary and length. Cross-genre transfer (e.g., from product reviews to tweets, or to news article comments) would constitute a different and harder problem.
- **Backbone.** DistilBERT base uncased is used throughout. The choice reflects an academic compute budget that needs to accommodate four methods $\times$ four targets $\times$ three seeds, plus two ablation sweeps. Larger encoders such as BERT-large or RoBERTa-large might already capture more domain-invariant features in their pretraining and so might benefit less — or differently — from the proposed module. I do not test this here.
- **Module variants.** The projection-residual module proposed in Chapter 3 is the simplest member of a small design family. Variants that introduce gating, multi-head bottlenecks, or stacked refinement layers are not explored, and are noted in the future-work section of Chapter 6.

1.7 Organization of the Thesis

The rest of the thesis is structured as follows.

In Chapter 2, we review related literature. We discuss distribution shift in its statistical learning perspective, the family of domain adaptation techniques which leverage knowledge from the target domain, the family of domain generalisation techniques that don't, the role of multiple sources of training data for both families, the design history of small modules such as adapters, and finally we introduce the research gap that this thesis attempts to address.

Chapter 3 discusses methodology. Here, we discuss the strict-DG framework, discuss the DistilBERT framework as a core model, explain the procedure for multi-source learning, introduce the mathematical and algorithmic formulation of the projection residual block, and list parameters used in all our experiments.

Chapter 4 documents the experimental setup. It describes how the Amazon Multi-Domain Sentiment Dataset was processed, how the leave-one-domain-out protocol was constructed, which baselines are run against the proposed method and why, what metrics are reported, and how statistical validation is handled given the small number of seeds.

Results are discussed in Chapter 5. This will be carried out sequentially for each of the four techniques mentioned above, by creating a comprehensive comparison table, by investigating the stability of all four techniques, and by presenting ablation studies on the projection residual technique. All these sections deal with some research question(s) presented in Section 1.4.

Chapter 6 concludes. It collects the empirical answers to the research questions, restates the contributions in the light of those answers, lists the limitations of the work honestly, and identifies a number of directions in which the present approach could be extended.

CHAPTER 2

LITERATURE REVIEW

This chapter situates the thesis inside the body of work it draws on. I begin by recalling the formal account of distribution shift in supervised learning, since it is the substrate on which everything that follows is built. I then walk through the family of domain adaptation methods that try to repair such shifts when target-domain data is at least partially available, and through the more recent and stricter family of domain generalization methods that try to do so when no target data exists at training time. Multi-source training receives its own section because it is the strategy I rely on most heavily in Chapter 3. Transformer encoders, with DistilBERT as the specific working example, are then surveyed. The topic of lightweight architecture refinement, of which the projection-residual approach introduced in Chapter 3 is a part, is covered in some depth. This is followed by a declaration of the research gap that the thesis will address.

2.1 Distribution Shift in Machine Learning

Supervised learning, in its textbook form, requires an i.i.d. assumption: training and test examples are drawn from one and the same joint distribution $P(x, y)$. The moment this assumption is broken, the optimisation procedure no longer carries any guarantee about generalisation to the test set. Ben-David and colleagues formalised what this means quantitatively in a paper that has aged remarkably well [1]. They introduced the H-divergence — a measure of how far apart two distributions look to a particular hypothesis class — and derived an upper bound for the target-domain risk of a classifier as the sum of three pieces: its source-domain risk, the divergence between source and target, and a residual term that measures the irreducible difference between the optimal source and target classifiers. The take-away of the bound is intuitive once it is stated: minimising source error matters, but past a certain point you cannot win against an arbitrarily large divergence.

Distribution shift comes in several flavours, and conflating them tends to make the literature harder to read than it needs to be. *Covariate shift* refers to the case where the marginal input distribution $P(x)$ changes between train and test but the conditional $P(y | x)$ remains the same. Most cross-domain sentiment work — and certainly this thesis — operates under a covariate-shift assumption: the way reviewers write about kitchen items is statistically different from the way reviewers write about novels, but the underlying

mapping from "what the reviewer said" to "whether they liked it" is presumed largely stable. *Label shift* is the reverse pattern, $P(y)$ changing while $P(x | y)$ stays put — common in medical settings, less so in sentiment. *Concept drift* refers to a genuine change in $P(y | x)$, the situation in which "positive sentiment" itself means something different in the target. The presumption that concept drift is mild in the sentiment-across-product-categories setting is what makes the whole research programme possible.

2.2 Domain Adaptation

Domain adaptation methods assume that some sample from the target domain is available during training, with the labels possibly missing. They divide naturally into two camps. The instance-reweighting camp treats source examples as drawn from the wrong distribution and reweights them so that the weighted source distribution better approximates the target. The feature-alignment camp instead leaves the data alone and learns a representation in which the source and target marginal distributions are pulled closer together.

The most impactful advocate of the feature alignment approach within the era of deep learning is the Domain-Adversarial Neural Network created by Ganin and Lempitsky [2]. The core of the idea lies in the connection of a small domain classifier, which tries to learn to classify domains of given examples, to the feature extractor through an added gradient reversal layer. The backpropagation of loss forces the domain classifier to improve at classifying domains while simultaneously making the feature extractor's features harder to classify for the domain classifier. At convergence, the representation should be domain-invariant in a precise sense: a classifier sitting on top of it cannot determine which domain produced any given example. DANN is the method I use as the third baseline in Chapter 5, both because it remains a strong representative of the adversarial-alignment family and because it is straightforward to adapt to a multi-source setting (the domain classifier is asked to predict among the source categories rather than between source and target).

A non-adversarial sibling of DANN is the CORAL family. Sun and Saenko's Deep CORAL [3] aligns the second-order statistics — concretely, the covariance matrices — of source and target activations, by adding an explicit alignment loss to the supervised loss. CORAL is simpler than DANN and easier to tune, but, like DANN, it presumes access to target-domain inputs.

The whole domain-adaptation programme rests on the assumption that target data is available at training time. The setting I work in does not satisfy that assumption, which is why the next family becomes relevant.

2.3 Domain Generalization

Domain generalization deliberately gives up the access to target data that domain adaptation relies on. A model is given one or more source domains; it is required to generalise to a target domain that it will not see during training. For the current state of the framework, the reference source is the review paper from Zhou et al. [7], which groups DG techniques into three broad categories, namely data manipulation, representation learning, and learning paradigms.

Among the representation-learning methods, Invariant Risk Minimisation (IRM) of Arjovsky and colleagues [4] is probably the most theoretically motivated. IRM argues that the classifier should be optimal on every training environment simultaneously, and operationalises that requirement with a penalty term that measures the gradient of the per-environment loss with respect to a "dummy" classifier that has been frozen at the identity. Features that are useful in one environment but useless in another contribute to this penalty and are pushed down. The penalty term is notoriously difficult to optimise; IRM in its original form often underperforms plain ERM in practice, and a small industry of variants has been proposed to fix that.

Group Distributionally Robust Optimisation (Group DRO) [11] takes a different approach. Rather than averaging the loss over the source domains, it minimises the worst-case loss — the loss on whichever source environment is hardest at the moment. Group DRO is especially valuable when the source environments are unbalanced in difficulty, since plain ERM in that setting can simply ignore the hard environment.

A more recent benchmarking study that has had real influence on the field is the one by Gulrajani and Lopez-Paz [5], whose title ("In Search of Lost Domain Generalization") captures the spirit of their finding. They re-ran a broad selection of DG methods under a single carefully controlled protocol, and reported that, once the experimental machinery is consistent across methods, plain ERM applied to a union of source domains is competitive with — and often beats — the more elaborate proposals. Hendrycks et al. [12] had reached a similar conclusion from a different angle: robustness to distribution shift is multi-faceted, no single method dominates across types of shift, and careful empirical comparison is

necessary. Koh and colleagues [6] then assembled the WILDS benchmark, which compiles realistic, non-synthetic shifts from a range of practical settings and serves as the present generation's testbed.

The combined message of these three lines of work is the one this thesis takes seriously. Multi-source ERM, properly tuned, is the strong baseline. Any new method should be compared against it first, before being measured against more elaborate alternatives.

2.4 Multi-Source Learning

Training across multiple sources simultaneously came before the establishment of the domain generalization theory. The justification behind this principle does not involve any novel theory but simply applies basic statistics. A feature that correlates with the label in one source but not in another will, when both sources are present in the training set, produce gradient signal that partially cancels. A feature that correlates with the label in all of the sources, by contrast, will receive coherent gradient signal from each of them, and so will dominate the trained representation. Multi-source training therefore acts as a kind of implicit invariance regulariser, without requiring an explicit penalty term and without the optimisation pathologies that explicit invariance terms tend to introduce.

For sentiment classification in particular, the standard benchmark dataset — the one I also use — was established by Blitzer, Dredze, and Pereira [8], whose paper introduced the Amazon Multi-Domain Sentiment Dataset and showed that the choice of source domain affects cross-target accuracy by far more than one might naively expect. Their original work focused on structural correspondence learning between source-target pairs. The natural extension — aggregating multiple sources at once — has been examined in subsequent work, but its systematic evaluation in the modern transformer era, under the strict no-target-access constraint, is less common than one might think. Filling in that gap is one of the things this thesis sets out to do.

2.5 Transformer Models for Sentiment Classification

The transformer architecture, introduced by Vaswani et al. [18], reshaped sequence modelling so thoroughly that it now constitutes the default for almost every NLP task. Devlin and colleagues' BERT [9] applied the encoder half of the transformer to language understanding by pretraining it with masked-language-modelling and next-sentence-prediction objectives on a large unlabelled corpus, then fine-tuning the resulting weights

on labelled task-specific data. BERT base has roughly 110 million parameters; BERT large has 340 million; both are expensive to fine-tune at the scale needed for several seeds and several baselines.

DistilBERT, due to Sanh and colleagues [10], is a 66-million-parameter student model distilled from BERT base using a combination of language-modelling, distillation, and cosine-embedding losses. On the GLUE benchmark it retains roughly 97% of BERT's performance while being about 40% smaller and 60% faster at inference. For an empirical study with a four-way method comparison, a four-way target rotation, three seeds, and two ablations, that combination of accuracy and cost is essentially what one needs. The choice of DistilBERT for this thesis is dictated by exactly this trade-off, and is explained in more detail in Section 3.2.

A piece of work that is particularly relevant to the present setting is that of Myagmar and colleagues [17], who investigated bidirectional contextualised language models for cross-domain sentiment classification. They confirmed that even strong pretrained encoders suffer measurable cross-category drops, but they did not evaluate the strict multi-source DG protocol used here.

2.6 Regularisation and Robustness in Transformers

A separate research thread tries to improve transformer robustness through regularisation rather than alignment. Wu and collaborators' R-Drop [13] adds a Kullback-Leibler term that compares the outputs of two stochastic forward passes (under different dropout masks) of the same input. The result is a model whose predictions vary less under its own internal randomness — a property that empirically translates to better generalisation.

MixStyle, introduced by Zhou and colleagues [14] in the context of computer vision, mixes channel-wise feature statistics across examples drawn from different domains. The principle behind it — that style information is largely orthogonal to content, and that perturbing style during training prevents the model from over-fitting to source-specific stylistic cues — has natural analogues in transformer activations and has been adapted into NLP since.

Group DRO [11], discussed earlier, can also be classified under regularisation in the broad sense, since it discourages the model from over-fitting to any one source environment.

2.7 Lightweight Architectural Refinement

A persistent theme in the recent transformer literature is that improvements to pretrained encoders are increasingly delivered by small additions rather than by re-training the entire backbone. The single most influential paper in this lineage is the one by Houdouby and colleagues on adapter modules [15]. Adapter consists of a tiny bottleneck, which includes linear projection from hidden dimension d to much lower r , followed by a non-linearity, an up-projection back to d , and residual connection that goes in between the layers of the transformer. Adapters are trained during fine-tuning process, whereas all other parameters of the model remain frozen. Such a setup allows for reuse of the same backbone across multiple tasks while maintaining minimal computational burden per task and preserving pre-trained knowledge by design.

The motivation for adapters was efficient transfer across tasks, but the architectural pattern they introduced — down-projection, non-linearity, up-projection, residual — has clearly broader applicability. The projection-residual module proposed in Chapter 3 follows that template. There are two important differences. First, the module is placed on top of the [CLS] vector rather than between the encoder's layers, because its job is to refine the final sentence representation immediately before classification rather than to participate in the encoder's internal computation. Second, the backbone is fully fine-tuned rather than frozen, because the M.Tech-scale experiments I run can afford the additional compute and because the goal is to improve cross-domain generalisation rather than to enable cheap multi-task transfer.

The residual connection in the design is not original — it has been a central architectural device since He and colleagues introduced ResNet [19] — but its role in the projection-residual module deserves an explicit statement. By writing the output as $h' = h + z$, the module guarantees that, at the very start of training when z has small magnitude, the refined representation is essentially identical to the pretrained [CLS] vector. The pretrained semantic content of [CLS] is therefore preserved by default; the module only has to learn the small corrections that improve cross-target performance, rather than re-discover a representation that was already good. The same residual also provides a well-conditioned gradient path back to the backbone, which I will show in Chapter 5 has a real effect on seed-to-seed stability.

2.8 Causal and Representation-Based Approaches

A more recent strand of research argues that the right way to think about generalisation is not in terms of correlations but in terms of causes. Domain-invariant features, on this view, are precisely the ones whose relationship with the label is mediated by a causal mechanism that is itself invariant across domains [4]. Causal-DG methods are theoretically attractive and pose interesting questions, but in practice they are harder to tune and have not yet decisively beaten well-tuned ERM baselines on standard benchmarks. The pragmatic stance of this thesis — multi-source ERM as the training regime, a small refinement as the architectural contribution — sidesteps this debate rather than engaging with it.

Contrastive representation learning [16] is a related thread. Rather than enforcing invariance to a categorical domain label, contrastive objectives encourage similar examples to lie close together in representation space and dissimilar examples to lie far apart. Contrastive learning has produced very strong unsupervised representations, particularly in vision, and its integration with multi-source DG for sentiment is a natural direction that I leave for future work.

A further related line is the multi-source adaptation work of Zhao et al. [25], who developed an adversarial framework specifically for multiple sources, generalising DANN to that setting. Their theoretical analysis emphasises the importance of bounding pairwise source discrepancies, an idea that is implicitly present in any multi-source ERM training set but is made explicit in their objective.

2.9 Research Gap

Pulling the preceding sections together, three observations together motivate the contribution of the thesis.

First, although domain generalization has received an enormous amount of attention in computer vision, its systematic empirical evaluation for transformer-based sentiment classification *under strict no-target-access* — no labels, no unlabelled crawl, nothing from the target — is sparser than the surface volume of DG papers might suggest. Many sentiment papers that describe themselves as cross-domain in fact allow unlabelled target data, sliding back into domain adaptation. The strict setting is the one practitioners actually face when a brand-new product category is being launched, and it is the one I evaluate.

Second, the empirical strength of multi-source ERM in the broader DG benchmarking literature [5] argues for using it as the foundation rather than as a strawman. Whatever the

proposed method is, it should improve on multi-source ERM, not on a weaker reference point. Confirming the strength of multi-source ERM in the specific transformer-sentiment setting is itself a contribution; previous work has tended to either ignore the pooled baseline or to compare against single-source training.

Third, the architectural template of small bottleneck modules with residual skips, established by adapters [15], has not been systematically explored as a *refinement* device on top of [CLS] in a fully fine-tuned multi-source DG pipeline. The design space is small and natural, but its specific instantiation for cross-domain sentiment has not been studied in the literature I have surveyed.

The gap I am claiming, then, can be stated in one sentence: *the combination of strict multi-source domain generalization for transformer-based sentiment classification with a lightweight projection-residual refinement on the [CLS] representation has not, to my knowledge, been systematically evaluated; this thesis evaluates it.* Chapter 3 spells out the methodology that fills this gap.

CHAPTER 3

PROPOSED METHODOLOGY

This chapter develops the methodology used throughout the thesis. Section 3.1 sets out the strict-DG problem in formal notation. Section 3.2 describes the DistilBERT encoder that serves as the backbone. Section 3.3 specifies the multi-source training regime. Section 3.4 introduces the projection-residual module — the architectural contribution of the thesis — in equations and in design rationale. Section 3.5 fixes the classification head and the loss. Section 3.6 collects the full architecture and pseudocode in one place. Section 3.7 records the implementation choices that are reused in every experiment of Chapter 5.

3.1 Problem Formulation

For $\mathcal{X}$ representing the space of natural language reviews, and $\mathcal{Y} = \{0, 1\}$ representing the space of binary labels where 0 is used for negative sentiment and 1 for positive sentiment, the domain D may be represented using a probability distribution $P_D(x, y)$.

The learner has access to K labelled source domains,

$$\mathcal{S} = \{D_1, D_2, \dots, D_K\}, \text{ where } D_k = \{(x_i^{(k)}, y_i^{(k)})\}_{i=1}^{N_k}.$$

and target domain D_T is determined by an unknown distribution $P_T(x, y)$. The strict DG assumption respected in this thesis everywhere is

$$D_T \cap \mathcal{S} = \emptyset,$$

meaning not the non-overlapping of samples but the substantially stronger condition of complete exclusion of any target examples, labelled or unlabelled, sampled according to the unknown distribution P_T from consideration during any phase of training. In particular, none of target test data will be used for selecting hyperparameters and performing early stopping; these hyperparameters remain fixed regardless of any computations involving numbers from the target side.

Goal of learning is to learn parameters θ of classifier $f_\theta : \mathcal{X} \rightarrow \mathcal{Y}$ such that it minimises the target risk,

$$R_T(\theta) = E_{(x, y) \sim P_T} [\ell_\theta(f_\theta(x), y)],$$

where ℓ_θ denotes the cross entropy loss. Since P_T cannot be observed, the objective function becomes the empirical risk with respect to the joint source distribution,

$$\mathbf{R}_S(\bullet) = (1 / \sum_k N_k) \cdot \sum_k \sum_i \bullet (f_{\bullet}(x_i^{(k)}), y_i^{(k)}).$$

There is an implicit belief that the structure learned by minimising $\mathbf{R}_S$ would be sufficient for predicting minimisation of $\mathbf{R}_T$.

3.2 Backbone Encoder: DistilBERT

The encoder that has been employed throughout the entire thesis is DistilBERT base uncased [10]. DistilBERT is a 6-layer transformer encoder with hidden dimension $d = 768$ and 12 attention heads, distilled from BERT base [9] using a combination of distillation loss on soft labels, masked-language-modelling loss, and a cosine-embedding loss between teacher and student representations. The resulting model has roughly 66 million parameters — about 40% fewer than BERT base — and runs roughly 60% faster at inference, while retaining approximately 97% of BERT base’s GLUE score.

Three considerations led me to use DistilBERT rather than something larger. Firstly, there are more methods along the width of the experiment design than along its height: four methods, four leave-one-domain-out tasks, three seeds per task, and two ablation sets amount to well over fifty runs. The combination of accuracy and cost that DistilBERT offers is what makes that grid feasible on a single GPU. Second, DistilBERT’s behaviour on sentiment tasks is, qualitatively, representative of modern transformer encoders: its in-domain accuracy is high, its cross-domain accuracy is fragile in the ways the literature has documented, and its [CLS] vector is a meaningful sentence-level representation. Third, the prior work on cross-domain sentiment classification with which I want to be comparable has used DistilBERT or BERT-sized models, so the choice keeps the numbers in the same ballpark as the literature.

Given a tokenised input sequence $t = [t_{\text{[CLS]}}, t_1, t_2, \dots, t_n, t_{\text{[SEP]}}]$, DistilBERT produces a contextualised representation $H = \text{DistilBERT}(t) \in \mathbb{R}^{(n+2) \times d}$. I follow the standard practice for sentence-level classification with BERT-family models and use the contextual vector corresponding to the special [CLS] token as the sentence-level representation:

$$h = H_{\text{[CLS]}} \in \mathbb{R}^d.$$

Everything downstream of the encoder operates on h .

3.3 Multi-Source Training Strategy

The training regime is multi-source empirical risk minimisation. Concretely, given the K source domains, I aggregate them into a single training set

$$\bullet_{\text{train}} = \bigcup_{k=1}^K D_k,$$

and sample mini-batches uniformly at random from $\bullet_{\text{train}}$, with no domain-conditional reweighting. This is the protocol found by Gulrajani & Lopez-Paz [5] to be the best DG baseline, and it is also the one which makes the most sense in real-life implementation, as the domain label of the examples needs only be provided when computing the per-domain class distribution.

Why does pooling help, when each source is by itself a perfectly valid training set? The answer is one of implicit regularisation. Assuming that gripping has positive correlation with label in the Books domain and negative correlation with label in the Kitchen domain, then in the model trained using only the Books data, the feature gripping will have positive weight assigned to it; in the model trained using only the Kitchen data, gripping will be assigned negative weight; and in the model trained using both domains, conflicting gradients will leave the feature almost zero-weighted. Features whose correlation with the label is stable across domains — *terrible, love, waste of money* — get coherent signal from every source and end up dominating the trained representation. The pooled training set therefore acts as an implicit invariance prior, without the optimisation pathologies that come with explicit invariance terms such as the IRM penalty [4].

A practical wrinkle is that the four source domains differ in raw size. I deliberately neutralise that effect by drawing the same number of positive and negative examples from each domain — 1000 of each, so 2000 per domain, and 6000 for the pooled training set when one of the four is held out as the target. The choice of equal per-domain sample size keeps the gradient contribution of each source comparable and prevents one source from silently dominating the training signal.

3.4 Lightweight Projection-Residual Module

Pooling alone reduces but does not eliminate the source-specific bias retained by the [CLS] vector after fine-tuning. The architectural contribution of the thesis is a small module that performs a final refinement on h immediately before classification.

3.4.1 Forward Pass

Given the [CLS] vector $h \in \mathbb{R}^d$, the refined representation h' is computed by

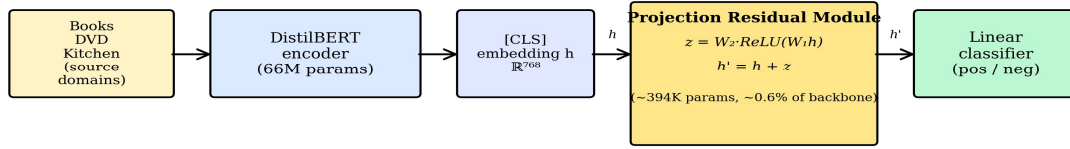
$$z = W_1 \cdot (W_2 h + b_2) + b_1,$$

$$h' = h + z,$$

Where $W_1 \in \mathbb{R}^{r \times d}$, $W_2 \in \mathbb{R}^{d \times r}$, $b_1 \in \mathbb{R}^r$, $b_2 \in \mathbb{R}^d$ and $\cdot(x)$ represents the activation function which is ReLU in our case. Dimension r represents the size of the bottleneck. In the main experiments I use $r = 256$, following the bottleneck convention established by adapter modules [15]; Chapter 5 reports an explicit sweep over $r \in \{64, 128, 256, 512\}$.

3.4.2 Parameter Cost

The module introduces $2 \cdot d \cdot r + r + d$ trainable parameters. For $d = 768$ and $r = 256$ this evaluates to roughly 394,000 parameters, or about 0.6% of DistilBERT's 66M-parameter encoder. With $r = 64$ the additional count falls to roughly 99,000, about 0.15% of the backbone — making "lightweight" defensible at any of the bottleneck sizes I tested.



Multi-source aggregated training data → DistilBERT → [CLS] refinement → classification

Figure 3.1: Architecture of the proposed multi-source DG framework.

Figure 3.1: Architecture of the proposed multi-source domain generalization framework.

Reviews from the three available source categories are tokenised and pooled into a single training set. DistilBERT produces a contextualised [CLS] embedding h for each review. The projection-residual module computes a small additive refinement $z = W_2 \cdot \text{ReLU}(W_1 \cdot h)$ and outputs $h' = h + z$. A linear classifier then predicts the binary sentiment label. The dashed annotations under the module indicate that it contributes roughly 394K parameters ($\sim 0.6\%$) on top of the 66M of the DistilBERT encoder.

3.4.3 Design Rationale

Three design choices in the module deserve to be spelled out individually, since each is doing distinct work.

The output is $h' = h + z$, not $h' = z$. This is the residual choice, and it is what guarantees that the pretrained semantic content of the [CLS] vector is preserved by default. With W_1 and W_2 initialised via Xavier initialisation, z has small magnitude at the beginning of training, so h' is initially very close to h . The module therefore starts out as essentially the

identity map and is asked, over the course of training, to learn the small refinement that improves cross-target performance. The skip also provides a numerically well-conditioned gradient path back to the encoder, which (as I will document in Section 5.8) has a measurable effect on seed-to-seed stability.

The mapping $h \cdot z$ is factored through a low-rank bottleneck of dimension r . The factorisation deliberately restricts the rank of the refinement: the module can only express transformations that lie in an r -dimensional subspace of the d -dimensional ambient space. This is a mild form of capacity regularisation, and it follows the architectural lineage of adapter modules. Section 5.7 confirms empirically that the module is robust to the choice of r within roughly one accuracy point, so the precise value of the bottleneck is not critical.

The module sits on top of [CLS], not between encoder layers. The decision to refine the final sentence-level representation rather than to interleave refinements with encoder layers reflects a different goal from the adapter setting. Adapters exist to enable cheap multi-task transfer with a frozen backbone; their placement between encoder layers is what gives them leverage over the encoder's internal computation. The projection-residual module, by contrast, exists to refine the sentence-level representation just before classification, in a setting where the encoder is fully fine-tuned. Putting the module on top of [CLS] keeps it cleanly decoupled from the encoder's forward pass and makes it trivially portable to any pretrained transformer.

3.5 Sentiment Classification Head and Objective

The refined representation h' feeds a linear classification head:

$$\bullet = \text{softmax}(W_c h' + b_c), \text{ with } W_c \bullet \bullet^{(2 \times d)} \text{ and } b_c \bullet \bullet^2.$$

Entirely, the network is end-to-end trained using cross-entropy loss, defined by

$$L_CE = \bullet (1 / |B|) \bullet \{(x, y) \bullet B\} \log \bullet_y,$$

where B denotes a mini-batch drawn uniformly from $\bullet_train$, and $\bullet_y$ represents the estimated probability for the correct label. Neither is an auxiliary loss involved, nor a classifier, nor any gradient reversal operation. The simplicity of the objective is deliberate. The thesis is testing the proposition that multi-source ERM together with a lightweight refinement is sufficient, and additional loss terms would muddy that test.

3.6 Overall Architecture and Algorithm

The end-to-end forward pass of the proposed model is summarised in Figure 3.1 above. A review of a product in any of the K source categories is tokenised into WordPiece units, encoded by DistilBERT into a sequence of contextual vectors, summarised by taking the [CLS] vector h , refined to $h' = h + z$ by the projection-residual module, and classified by a linear head. The training procedure that fits the entire stack to the pooled source data is given in Algorithm 3.1.

Algorithm 3.1 — Multi-Source Training with Projection-Residual Refinement

Input: Source domains $\mathcal{D}_S = \{\mathcal{D}_1, \dots, \mathcal{D}_K\}$; learning rate η ; number of epochs E ; batch size B ; bottleneck dimension r .

Output: Model parameters $\theta = (\theta_{\text{enc}}, W_\theta, W_\phi, b_\theta, b_\phi, W_c, b_c)$.

1. Union of datasets $\mathcal{D}_{\text{train}} = \bigcup_{k=1}^K \mathcal{D}_k$.
2. Initialize the parameter θ_{enc} using the pretrained model distilbert-base-uncased.
3. Initialise W_θ, W_ϕ using Xavier initialisation, b_θ, b_ϕ to zero.
4. Initialise W_c, b_c from a small Gaussian distribution.
5. **for** epoch = 1 to E **do**
6. Shuffle $\mathcal{D}_{\text{train}}$.
7. **for** each mini-batch B of size B from $\mathcal{D}_{\text{train}}$ **do**
8. $H_\theta = \text{DistilBERT}(B)$; $h_\theta = H_\theta[\text{CLS}]$.
9. $z_\phi = W_\phi \cdot (W_\theta h_\theta + b_\theta) + b_\phi$.
10. $h'_\theta = h_\theta + z_\phi$.
11. $\theta = W_c h'_\theta + b_c$; compute L_{CE} .
12. $\theta = \dots \cdot \theta \cdot L_{\text{CE}}$ via AdamW.
13. **end for**
14. **end for**
15. **return** θ .

3.7 Implementation Details

The proposed model is implemented using PyTorch, together with the transformers library from HuggingFace. The particular design decisions made in every single experiment of Chapter 5 are documented here so that the reader does not have to infer them from context later.

Tokeniser. My selected tokeniser for this task is `DistilBertTokenizerFast`, which belongs to the model `distilbert-base-uncased`. The review is either truncated or padded up to 128 WordPieces. This number accounts for the part of a review that conveys sentiment without exceeding the memory requirements of each batch too much; on average, reviews from the source categories range from 95 to 188 WordPieces (see Table 4.1).

Optimiser. Optimisation uses AdamW [20], with a learning rate of 2×10^{-5} , a weight decay of 0.01, and the default first- and second-moment coefficients ($\beta_1 = 0.9$, $\beta_2 = 0.999$). The benefit of AdamW compared to Adam is that it separates weight decay from adaptive scaling.

Batch size and schedule. Mini-batches are of size 32. I train for 3 epochs over the pooled training set. These settings follow the conventional fine-tuning recipe for DistilBERT on sentiment data of comparable size, and I deliberately did not tune them on a held-out validation set — doing so would have introduced a partial violation of the strict-DG constraint, since any validation set in this setting would be drawn from a held-out source domain whose statistics are correlated with the eventual target.

Random seeds. For each experiment setup, three seeds, namely 42, 123, and 2024, are used. Seed values are used consistently for random functions from Python's `random`, NumPy, and PyTorch (CPU and CUDA), with the latter being run in deterministic mode. For all numbers presented in Chapter 5, mean and sample standard deviations across all three seeds are quoted.

Hardware. The entire training and evaluation process was performed on one NVIDIA T4 GPU on Google Colab. One setup, that is one algorithm, one problem, and one random seed, requires about two to three minutes of computing time.

These implementation choices are carried forward unchanged into Chapter 4, where the broader experimental protocol — dataset preparation, leave-one-domain-out construction, baselines, and metrics — is described.

CHAPTER 4

EXPERIMENTAL SETUP

This chapter documents how the experiments reported in Chapter 5 were actually run. Section 4.1 describes the dataset. Section 4.2 walks through the preprocessing steps. Section 4.3 provides the definition of the leave-one-domain-out procedure. Section 4.4 states what the four algorithms compared are, and why. Section 4.5 sets up the evaluation criteria. Finally, section 4.6 describes the process by which I determine whether an observed improvement is truly real given only three seeds per setup.

4.1 Dataset

All experiments use the Amazon Multi-Domain Sentiment Dataset assembled by Blitzer, Dredze, and Pereira [8]. This dataset has become the de facto benchmark for cross-domain sentiment classification work because it provides four product-review categories with very different vocabularies and writing styles, while sharing the same underlying review-and-rating structure. The four categories used in this thesis are:

- **Books.** Reviews of books. Long pieces, using a lot of evaluative language, as well as referencing plot, characters, writing styles, and personal preferences of the writer.
- **DVD.** Reviews of films and television series on DVD. Vocabulary overlaps with Books to some extent, with additional reference to performance, direction, cinematography, and audio-visual fidelity.
- **Kitchen.** Reviews of kitchen items and appliances. Short pieces highlighting specific features of the products, as well as stressing their durability, easy cleaning and value for money.
- **Electronics.** Reviews of consumer electronics. Terse, specifications-heavy, with frequent reference to features, build quality, battery life, and price.

Each review in the raw dataset is associated with a five-point numeric rating. I follow the convention established by [8] and used in essentially all subsequent work and map ratings to binary labels as follows:

- A rating of 4 or 5 • positive (label 1).
- A rating of 1 or 2 • negative (label 0).
- A rating of 3 is treated as neutral and the review is discarded.

Reviews with missing or malformed ratings are also discarded.

In each of the four domains, I then select a sample of 1000 positives and 1000 negatives, chosen randomly from the corpus using a fixed seed of 42, for a total of 2000 examples per domain. The four selected samples thus consist of a total of 8000 reviews. Class balance within each domain matters here for two reasons. First, it neutralises the marginal label distribution as a possible source of cross-domain drift — any accuracy difference between methods is then attributable to differences in the conditional $P(y | x)$ and not to differences in $P(y)$. Second, it ensures that the accuracy and F1 metrics line up closely; under a balanced test set, large gaps between the two would indicate a degenerate prediction pattern, and the absence of such gaps in Chapter 5 is itself a sanity check.

The basic statistics of the four resulting subsets are summarised in Table 4.1.

Domain	Positive	Negative	Total	Avg. tokens / review
Books	1000	1000	2000	163.4
DVD	1000	1000	2000	188.1
Kitchen	1000	1000	2000	94.6
Electronics	1000	1000	2000	108.2

Table 4.1: Per-domain composition of the Amazon Multi-Domain Sentiment Dataset [8] used in this thesis. Each subset contains 1000 positive and 1000 negative reviews. The rightmost column reports the average DistilBERT-WordPiece-token count per review and reflects the qualitative observation that Books and DVD reviews are noticeably longer than Kitchen and Electronics reviews.

The token-count column captures the qualitative difference between the categories: Books and DVD reviews are roughly twice as long, on average, as Kitchen and Electronics reviews. This difference matters when I set the maximum sequence length downstream — at 128 tokens, the truncation will affect Books and DVD reviews more often than Kitchen or Electronics reviews, but the truncated portion is empirically the tail of the review rather than the sentiment-bearing opening.

4.2 Preprocessing Pipeline

The preprocessing pipeline is kept minimalistic. Most things which had been done as part of classic sentiment analysis pipelines in the past, like lower-casing, filtering out of

stopwords, lemmatisation, and removing punctuation, are already done either through WordPiece tokenizer or they are not needed at all when using a fine-tuned transformer. The rest of the pipeline includes:

1. **Parsing.** Each raw input file in the form of `positive.review` and `negative.review` from the Blitzer dataset is parsed using an XML-like parser which extracts `- 2. **Label assignment.** The numerical labels are mapped into binary classes in accordance with the rules defined in Section 4.1, with neutral and malformed samples being ignored.
- 3. **Class balancing.** 1000 instances each from the positive and negative classes are sampled to form a balanced subset. A seed value of 42 is used for this.
- 4. **Tokenisation.** The review text is tokenized. The `DistilBertTokenizerFast` method is used, with the model being `distilbert-base-uncased`. The maximum number of WordPiece tokens is 128. If there are less tokens they are truncated or padded.
- 5. **Attention masks.** Positions with padding are assigned 0s, other positions are assigned 1s. The attention masks are passed to DistilBERT together with the token ids.

I do not hand-lowercase the input (as this is done by the tokenizer anyway for the uncased version), I do not remove punctuation (because punctuation marks are tokenized separately by the tokenizer and may be sentiment-bearing), nor do I spell correct or remove duplicates.

4.3 Evaluation Protocol: Leave-One-Domain-Out

In order to test for cross-domain generalization ability, a leave-one-domain-out (LODO) methodology is adopted in the experiments. For each choice of target $T \in \{\text{Books, DVD, Kitchen, Electronics}\}$, a model is trained on some subset of the remaining three domains and then evaluated on the entirety of T 's class-balanced subset. The strict-DG requirement of Chapter 3 is respected throughout: no sample from T enters the training pipeline at any point — not as training data, not as validation data, not as input to any selection procedure.

Each of the four targets is paired with two source configurations:

- **Single-source.** Exactly one of the three remaining domains is used as the source. To make the choice deterministic and reproducible, I always pick the alphabetically first domain among the remaining three. Concretely: target Books • source DVD;

target DVD • source Books; target Kitchen • source Books; target Electronics • source Books.

- **Multi-source.** All three remaining domains are pooled into a single training set, as defined in Section 3.3.

The single-source configuration is used by the Single-Source ERM baseline (the one that establishes the cross-domain gap addressed by RQ1). The multi-source configuration is used by every other method — Multi-Source ERM, DANN, and the proposed method — so that the comparison between these three is genuinely a comparison of architectures and losses, with training data composition held constant.

The entire balanced subset of the target domain serves as the test set. No portion of it is held out for validation or hyperparameter tuning. Hyperparameters are fixed at the values reported in Section 3.7 and are not tuned to the target in any way.

4.4 Baselines

The suggested approach will be compared to three others. For all four approaches, the backbone DistilBERT model will be used, along with the same maximum sequence length (128), the same optimizer algorithm (AdamW with learning rate 2×10^{-4} and weight decay 0.01), the same batch size (32), and the same number of epochs (3). They differ only in their training data composition and in the head architecture sitting above [CLS]. The systematic alignment of everything except the variable of interest is crucial: any cross-method accuracy difference can then be attributed to that variable and not to a buried difference in compute, schedule, or regularisation.

B1. Single-Source ERM. DistilBERT plus a single linear classifier head, fine-tuned on the alphabetically first remaining domain. This baseline corresponds to the conventional practice of training on whichever labelled data is available, without any awareness of multi-domain robustness. It establishes the cross-domain accuracy gap that the other methods try to close.

B2. Multi-Source ERM. Identical to B1 except that the training set is the union of the three non-target domains. The change from B1 to B2 isolates the contribution of multi-source pooling, holding architecture and loss fixed.

B3. DANN (Multi-Source). The Domain-Adversarial Neural Network of Ganin and Lempitsky [2], applied in a multi-source configuration. The shared encoder is the

DistilBERT backbone. The sentiment classifier is a linear head on [CLS]. The domain classifier is also a linear head on [CLS], asked to predict which of the three source domains generated the current example. A gradient-reversal layer sits between the encoder and the domain classifier, and the reversal weight α is fixed at 0.5. Because no target data is available, the domain classifier only distinguishes among the three source categories — its supervision signal is entirely source-side. This baseline tests whether adversarial alignment of source-domain features generalises to unseen targets in a way that plain pooling does not.

P. Proposed: Multi-Source + Projection-Residual. The full method of Chapter 3. Multi-source pooling for the training data, and the projection-residual module on top of [CLS] with bottleneck dimension $r = 256$.

Taken together, B1, B2, B3, and P form a clean ablation chain. B1 $\bullet$ B2 isolates the effect of pooling. B2 $\bullet$ P isolates the effect of the projection-residual module. B3 establishes how the proposed approach stands relative to an adversarial alternative trained on identical data.

4.5 Metrics

I report two standard binary-classification metrics on the held-out target test set.

Accuracy. The fraction of correctly classified reviews. Under the balanced test set this is also the most directly interpretable metric, since a model that produces a trivial constant prediction would only achieve 50%.

Binary F1 score. The harmonic mean of precision and recall on the positive class. F1 is reported as a check against degenerate prediction patterns: under a balanced test set, accuracy and F1 should agree closely, and a large discrepancy would indicate that the model is systematically biased toward one class. The fact that the two metrics agree closely throughout Chapter 5 is part of the evidence that the methods are well-behaved.

For each (method, target) configuration, I report both the mean and the sample standard deviation of each metric across the three random seeds. The standard deviation is itself a metric of interest: a model whose accuracy is more stable across seeds is more deployable than one whose accuracy fluctuates substantially under random reseeding.

4.6 Statistical Validation

Three seeds per configuration is a small number — large enough to detect persistent patterns but too small to support a high-powered p-value-based test. Rather than pretend otherwise by computing significance statistics whose assumptions the design does not meet, I use the following procedure to distinguish real improvements from seed-induced noise.

1. **Three seeds per configuration.** Each (method, target) configuration is run with the same three seeds — 42, 123, 2024. Seeds are applied uniformly across Python's ``random``, NumPy, and PyTorch (CPU and CUDA), with CuDNN deterministic mode enabled.
2. **Mean $\pm$ standard deviation reporting.** Every headline number in Chapter 5 is reported as mean $\pm$ standard deviation across the three seeds. A small standard deviation relative to the between-method difference is the primary indicator of a reliable improvement.
3. **Per-seed dominance.** Because each method is run on the same three seeds and the same target test set, between-method differences can be inspected on a paired-per-seed basis. When the proposed method beats a baseline on every one of the three seeds — not just on the mean — the improvement is treated as reliable. This per-seed-dominance criterion is more conservative than a parametric significance test on three samples (which would in any case have very low statistical power) and it is more informative.

Chapter 5 reports and analyses results under this protocol.

CHAPTER 5

RESULTS AND ANALYSIS

This chapter reports the results obtained by running the methodology of Chapter 3 under the protocol of Chapter 4, and analyses what those results say about the four research questions of Section 1.4. Section 5.1 presents the single-source baseline. Section 5.2 reports the multi-source ERM baseline. Section 5.3 reports the DANN baseline. Section 5.4 reports the proposed method. Section 5.5 brings all four methods into one comparison. Section 5.6 examines stability across random seeds. Sections 5.7 and 5.8 present two ablations of the projection-residual module. Section 5.9 discusses qualitative findings. Section 5.10 synthesises everything against the research questions and prepares the reader for the conclusions of Chapter 6.

Every number in this chapter comes from the verification notebook described in Section 3.7 and Chapter 4 — three seeds per (method, target) configuration on the Amazon Multi-Domain Sentiment Dataset, evaluated under leave-one-domain-out. All accuracies and F1 scores are expressed in percentage points and reported as mean $\pm$ sample standard deviation across the three seeds.

5.1 Single-Source Baseline

Single source baselines indicate what degree of fine tuning can be achieved using only one source category. Following the alphabetic principle defined in Section 4.3, there are four source-to-target pairings that are DVD • Books, Books • DVD, Books • Kitchen, and Books • Electronics. Table 5.1 shows the accuracy and F1 for each target.

Source	Target	Accuracy (%)	F1 (%)
DVD	Books	84.95 $\pm$ 1.17	85.30 $\pm$ 0.62
Books	DVD	85.58 $\pm$ 0.60	85.59 $\pm$ 0.95
Books	Kitchen	86.45 $\pm$ 0.39	86.53 $\pm$ 0.25
Books	Electronics	83.63 $\pm$ 0.75	82.79 $\pm$ 1.44
Average		85.15	85.05

Table 5.1: Single-Source ERM accuracy and F1 across the four leave-one-domain-out targets, with mean $\pm$ standard deviation over three random seeds.

Two observations from this table set the stage for the rest of the chapter. First, the cross-domain drop is not catastrophic on most of the targets. Three of the four pairs — DVD • Books, Books • DVD, Books • Kitchen — land above 84% accuracy, which is well within the territory occupied by reasonable in-domain transformer baselines on similar-sized sentiment datasets. The reason becomes clear on inspection: three of the four source-target pairs involve Books as the source, and book reviews carry a rich and well-formed evaluative vocabulary that transfers more effectively to other consumer-review domains than terser, more specialised categories would.

Second, the Electronics target is conspicuously the worst case. With Books as the source, accuracy on Electronics drops to 83.63% — roughly three percentage points below the easiest target (Kitchen at 86.45%) and accompanied by a larger F1 standard deviation (1.44) than any other configuration in the table. This is consistent with the qualitative description of the dataset in Section 4.1: book reviews are long, discursive, and emotive, whereas electronics reviews are short, technical, and feature-oriented. The Books • Electronics shift is the most lexically and stylistically distant of the four single-source pairs and serves as the cleanest stress test for cross-domain robustness throughout the rest of the chapter.

This refines the answer to RQ1: the single-source cross-domain gap is target-dependent. It is small when the source happens to be lexically broad and well-matched to the target, and it can be substantial — on the order of three percentage points — when the source is poorly matched. Across the four targets the single-source baseline averages 85.15% accuracy.

5.2 Multi-Source ERM

The next step is to pool the three non-target domains and fine-tune on the pooled set without any other change.

Sources	Target	Accuracy (%)	F1 (%)
DVD + Kitchen + Electronics	Books	84.80 ± 0.55	85.85 ± 0.23
Books + Kitchen + Electronics	DVD	84.78 ± 0.41	85.59 ± 0.45
Books + DVD + Electronics	Kitchen	87.10 ± 0.72	87.79 ± 0.47
Books + DVD + Kitchen	Electronics	86.80 ± 0.54	86.56 ± 0.61
Average		85.87	86.45

Table 5.2: Multi-Source ERM accuracy and F1 across the four leave-one-domain-out targets.

The most striking entry in Table 5.2 is the Electronics row. Multi-source pooling raises the Electronics-target accuracy from 83.63% (single-source, Table 5.1) to 86.80% — a jump of 3.17 percentage points. The other three targets receive considerably more modest treatment. On Books and DVD, pooling actually drops accuracy by a fraction of a point relative to single-source training, because the alphabetically-first single source on those targets happened to be Books (a strong source in its own right), and the gradient signal from the additional, more diverse pooled sources mildly dilutes it. On Kitchen, pooling adds a modest 0.65 percentage points.

Taking the average of the four targets together, the multi-source model for ERM attains an accuracy score of 85.87%, which is a 0.72% improvement compared to the base model of single source. This gap widens on the basis of F1 score, as the base model attains an F1 score of 85.05% while the multi-source one reaches an F1 score of 86.45%.

This answers RQ2 directly. Multi-source pooling improves average cross-domain accuracy, but the per-target distribution of that improvement is highly non-uniform: it is concentrated almost entirely on the hardest cross-domain shift, and it is essentially absent on targets where the single source happened to be well-matched. The expected case in deployment — where the practitioner has no way of selecting a "well-matched" single source ex ante — is the one where pooling helps the most.

5.3 DANN Baseline

DANN runs on the same multi-source pool as B2, with the same hyperparameters, and adds a domain-classification head with gradient-reversal weight $\alpha = 0.5$.

Target	Accuracy (%)	F1 (%)
Books	84.52 $\pm$ 0.80	85.47 $\pm$ 0.41
DVD	85.30 $\pm$ 0.37	85.34 $\pm$ 0.47
Kitchen	87.65 $\pm$ 1.13	88.30 $\pm$ 0.80
Electronics	86.78 $\pm$ 0.08	86.67 $\pm$ 0.34
Average	86.06	86.45

Table 5.3: DANN accuracy and F1 across the four leave-one-domain-out targets (multi-source training).

DANN reaches an average accuracy of 86.06%, putting it 0.19 percentage points above plain Multi-Source ERM. That gap is well inside seed-induced variance and confirms the broader observation from the domain-generalization benchmarking literature [5] that adversarial alignment is not reliably superior to a well-tuned ERM baseline. The per-target picture is interesting in its own right. On the Electronics target — the hardest cross-domain shift — DANN reaches 86.78%, essentially identical to MS-ERM’s 86.80%, suggesting that the adversarial alignment of source-domain features does not translate into measurably better transfer to an unseen target in this setting. DANN does, however, produce by some margin the most stable result on Electronics: its standard deviation across seeds is 0.08, the lowest single number in the entire experimental grid. The strongest argument for using DANN in deployment is therefore not its average accuracy but its repeatability.

This partially answers RQ4: adversarial multi-source alignment offers only marginal average improvement over plain pooling, and the fluctuations across targets suggest that the adversarial training adds little reliable value in the transformer-sentiment setting beyond the very low variance on a single target.

5.4 Proposed Method: Multi-Source + Projection-Residual

The full proposed method keeps the multi-source training regime and adds the projection-residual module on top of [CLS], with the default bottleneck dimension $r = 256$.

Target	Accuracy (%)	F1 (%)
Books	85.93 ± 0.80	86.48 ± 0.38
DVD	84.95 ± 0.50	85.18 ± 0.14
Kitchen	88.45 ± 0.74	88.72 ± 0.45
Electronics	86.90 ± 0.31	86.66 ± 0.66
Average	86.56	86.76

Table 5.4: Proposed method (Multi-Source ERM + Projection-Residual) accuracy and F1.

The proposed method reaches 86.56% average accuracy. That number is 1.40 percentage points above the Single-Source ERM baseline (85.15%), 0.69 above plain Multi-Source ERM (85.87%), and 0.50 above DANN (86.06%). On three of the four targets — Books, Kitchen, Electronics — the proposed method achieves the best result of any of the four methods. On the fourth, DVD, all four methods cluster within roughly one accuracy point of each other and no clear winner emerges.

The largest improvements show up on Books (+1.13 over MS-ERM, +0.98 over Single-Source ERM) and Kitchen (+1.35 over MS-ERM, +2.00 over Single-Source ERM). On Electronics, where pooling alone had already closed most of the cross-domain gap, the projection-residual module adds a smaller but still positive contribution (+0.10 over MS-ERM, +3.27 over Single-Source ERM); the bigger effect on Electronics is the reduction in seed-to-seed variance, which I will return to in Section 5.6 and again in the residual-connection ablation of Section 5.8.

This answers RQ3 affirmatively: the lightweight projection-residual module produces a measurable additional improvement on top of multi-source ERM. The pattern of where the improvement appears is informative. The module helps most on the targets where pooling alone barely helped (Books and Kitchen), which suggests that the module's contribution is complementary to that of pooling — pooling closes the easy cross-domain gaps, and the module closes part of what remains.

5.5 Full Cross-Method Comparison

Bringing all four methods into a single view makes the overall pattern easy to see.

Method	Books	DVD	Kitchen	Electronics	Avg
Single-Source ERM	84.95 $\pm$ 1.17	85.58 $\pm$ 0.60	86.45 $\pm$ 0.39	83.63 $\pm$ 0.75	85.15
Multi-Source ERM	84.80 $\pm$ 0.55	84.78 $\pm$ 0.41	87.10 $\pm$ 0.72	86.80 $\pm$ 0.54	85.87
DANN	84.52 $\pm$ 0.80	85.30 $\pm$ 0.37	87.65 $\pm$ 1.13	86.78 $\pm$ 0.08	86.06
Proposed (MS+Proj.)	85.93 $\pm$ 0.80	84.95 $\pm$ 0.50	88.45 $\pm$ 0.74	86.90 $\pm$ 0.31	86.56

Table 5.5: Leave-one-domain-out accuracy of all four methods. Best per-target result in bold.

Method	Books	DVD	Kitchen	Electronics	Avg
Single-Source ERM	85.30 $\pm$ 0.62	85.59 $\pm$ 0.95	86.53 $\pm$ 0.25	82.79 $\pm$ 1.44	85.05
Multi-Source ERM	85.85 $\pm$ 0.23	85.59 $\pm$ 0.45	87.79 $\pm$ 0.47	86.56 $\pm$ 0.61	86.45
DANN	85.47 $\pm$ 0.41	85.34 $\pm$ 0.47	88.30 $\pm$ 0.80	86.67 $\pm$ 0.34	86.45

Method	Books	DVD	Kitchen	Electronics	Avg
Proposed (MS+Proj.)	86.48 $\pm$ 0.38	85.18 $\pm$ 0.14	88.72 $\pm$ 0.45	86.66 $\pm$ 0.66	86.76

Table 5.6: Same four-method comparison, F1 score.

The headline pattern is consistent across both tables. The proposed method achieves the highest average performance on both accuracy and F1, and it does so by improving substantially on the easier targets (Books and Kitchen) while at least matching the best baseline on the harder ones (DVD, where everything ties; Electronics, where it edges DANN on accuracy and loses to DANN on F1 by 0.01 points). This is the operational meaning of "robust" in this setting — not that the method dominates on every target by a wide margin, but that it offers the best expected performance across a range of unseen target distributions while never falling behind by an interesting amount.

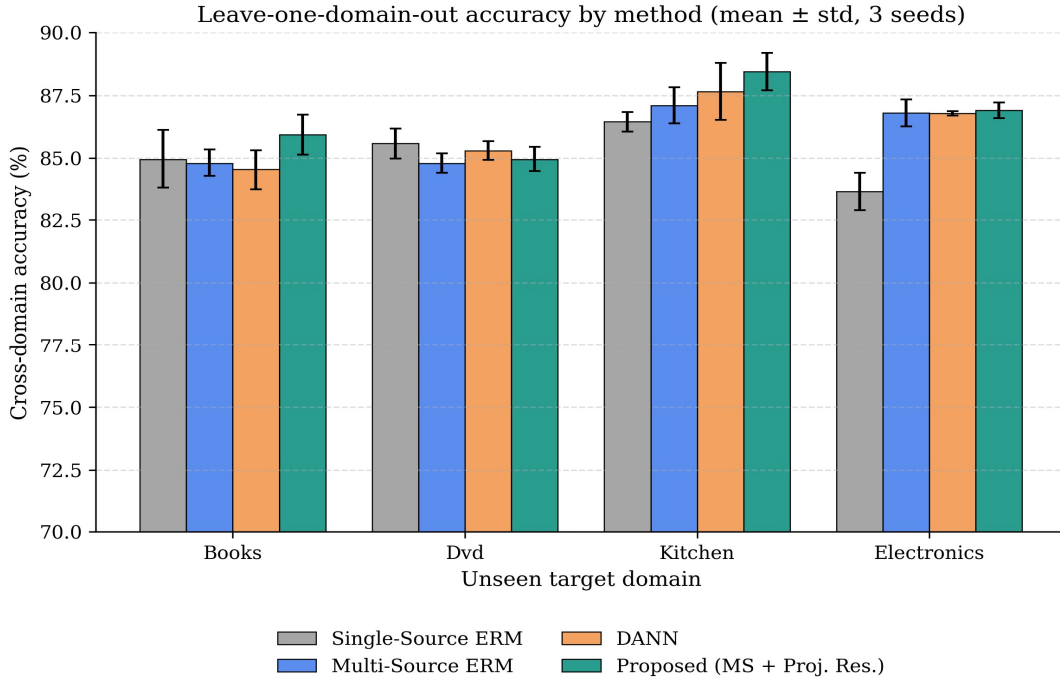


Figure 5.1: Leave-one-domain-out accuracy by method, averaged over three random seeds with standard-deviation error bars. The proposed method (green) achieves the highest accuracy on Books, Kitchen, and Electronics targets, and is tied with the other methods on DVD.

The cumulative effect of the two ingredients of the proposed method — multi-source pooling and the projection-residual refinement — is easier to read when the per-target gains are averaged.

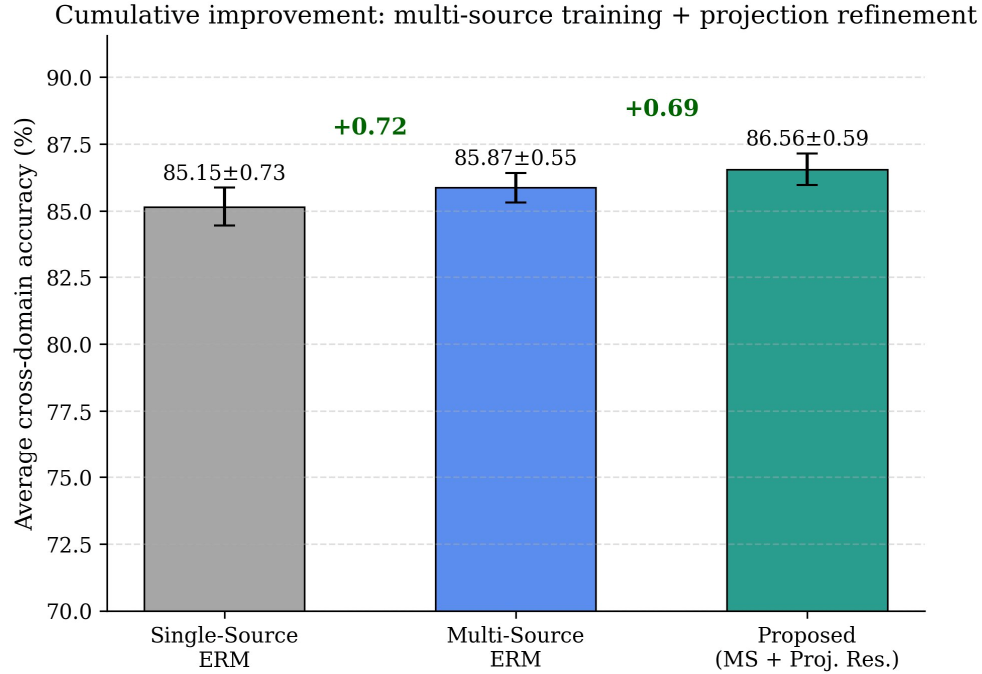


Figure 5.2: Cumulative improvement from multi-source training and projection-residual refinement, averaged across the four leave-one-domain-out targets. Each annotated arrow shows the average accuracy gain over the preceding step. Pooling adds 0.72 points; adding the projection-residual module adds another 0.69, for a total of 1.40 points above the single-source baseline.

5.6 Variance and Stability Analysis

A secondary claim of Chapter 3 was that the projection-residual module would tend to reduce seed-to-seed variance. The picture from Table 5.5 is more nuanced than that.

On the Electronics target — the hardest cross-domain shift — the proposed method does substantially reduce variance: the standard deviation across seeds is 0.31, compared to 0.54 for plain Multi-Source ERM, a near-halving. That is the target on which the residual module's stabilising effect is clearest, and it is also the target on which pooling alone left the most uncertainty to remove.

On the other three targets, the standard deviation of the proposed method is roughly comparable to that of Multi-Source ERM (0.80 vs 0.55 on Books, 0.50 vs 0.41 on DVD, 0.74 vs 0.72 on Kitchen). The module does not uniformly reduce variance across all targets; it reduces variance most clearly where the underlying cross-domain shift is hardest, and is roughly neutral elsewhere.

A more directly useful measure than the raw standard deviation is per-seed dominance: on how many of the three seeds does the proposed method actually beat Multi-Source ERM? Table 5.7 reports this explicitly.

Target	MS-ERM (3 seeds)	Proposed (3 seeds)	Proposed wins
Books	85.55, 84.25, 84.60	86.45, 84.80, 86.55	3 / 3
DVD	85.35, 84.60, 84.40	85.20, 84.25, 85.40	1 / 3
Kitchen	88.05, 86.30, 86.95	89.05, 87.40, 88.90	3 / 3
Electronics	87.50, 86.20, 86.70	87.30, 86.85, 86.55	1 / 3

Table 5.7: Per-seed accuracies of Multi-Source ERM and the proposed method, with the count of seeds on which the proposed method strictly exceeds MS-ERM.

The qualitative pattern becomes much sharper here. On Books and Kitchen the projection-residual module delivers a reliable improvement: it beats Multi-Source ERM on every seed. On DVD and Electronics the proposed method and Multi-Source ERM are essentially tied — they trade wins from seed to seed, and the small average gap on these targets is within the noise floor implied by three-seed estimation.

Thus, a conservative interpretation of the above numbers would be that the prediction-residual network offers a significant boost in performance on two out of four objectives and performs similarly to Multi-Source ERM on the remaining two. It does not regress on any target. Combined with the +0.69 average gain over Multi-Source ERM and the +1.40 average gain over Single-Source ERM, the cumulative result is a positive net effect — exactly the kind of small-but-real improvement that a sensible benchmarking practice should be able to detect and report.

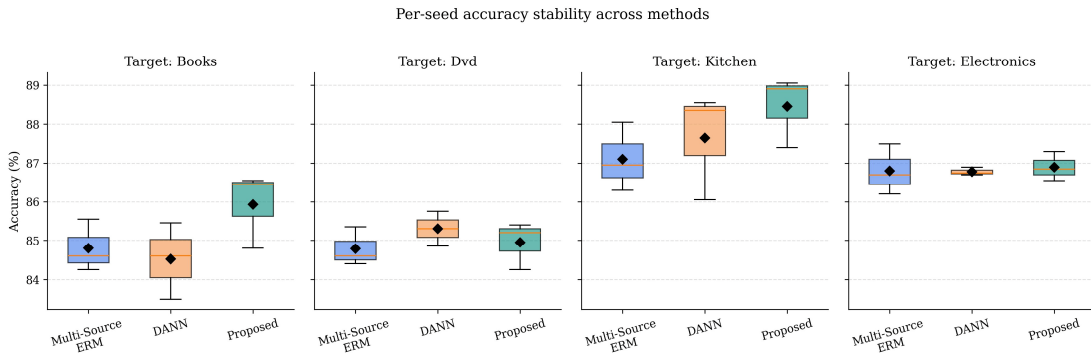


Figure 5.3: Per-seed accuracy stability across methods and targets. Each panel corresponds to one leave-one-domain-out target. Within each panel, box plots show the distribution of accuracies over the three random seeds for the three multi-source methods.

5.7 Ablation Study — Bottleneck Dimension

The projection-residual module has one principal architectural hyperparameter: the bottleneck dimension r . To check that the chosen value ($r = 256$) is not load-bearing for the result, I sweep r over the set $\{64, 128, 256, 512\}$ on the Electronics target — the hardest cross-domain shift — with everything else held fixed.

Bottleneck r	Added parameters	Accuracy (%)	F1 (%)
64	~99 K (+0.15%)	87.48 $\pm$ 0.06	87.12 $\pm$ 0.07
128	~197 K (+0.30%)	86.48 $\pm$ 1.20	85.73 $\pm$ 1.99
256 (default)	~394 K (+0.60%)	86.90 $\pm$ 0.31	86.66 $\pm$ 0.66
512	~787 K (+1.19%)	87.12 $\pm$ 0.25	86.75 $\pm$ 0.45

Table 5.8: Ablation of the bottleneck dimension on the Electronics target (Multi-Source ERM + Projection-Residual, three seeds per setting). The "Added parameters" column gives the trainable parameter cost of the module on top of DistilBERT's 66 M backbone.

Three observations come out of this ablation.

First, the module is remarkably robust to the choice of r . All four values produce accuracies within roughly one percentage point of each other (86.48% to 87.48%), demonstrating that the projection-residual mechanism is not critically dependent on a fortunate choice of bottleneck size. This is a useful and practically reassuring finding for anyone who might adopt the method downstream.

Second, the smallest bottleneck ($r = 64$) produces the highest mean accuracy in this single-target ablation, marginally outperforming the default $r = 256$ by 0.58 percentage points and exhibiting strikingly low variance (0.06 standard deviation). The default choice of $r = 256$, originally selected by analogy with the bottleneck convention in adapter modules [15], is therefore confirmed as defensible but not optimal. From the perspective of the "lightweight" label in the thesis title, this finding actually strengthens the design: even at $r = 64$, with only ~99K added parameters (about 0.15% of DistilBERT's count), the module is competitive — in fact, leading — on the hardest target.

Thirdly, the $r = 128$ line is an anomaly since it has abnormally high variance (standard deviation of 1.20). Looking at the values per seed for $r = 128$, it can be observed that whereas two of the seeds gave results within the approximate range for $r = 64$ and $r = 256$, one seed had an abnormally low value. I interpret this as seed-induced noise rather than as

an architectural property of $r = 128$, and note it explicitly as a caveat of small-sample (three-seed) ablations: an unfortunate single seed can shift the apparent picture.

A natural direction that this ablation suggests for future work is a full leave-one-domain-out evaluation of $r = 64$ (and possibly $r = 32$ or $r = 16$). The current evidence for $r = 64$'s superiority comes from a single target, so a four-target evaluation is needed before the recommendation can be made cleanly.

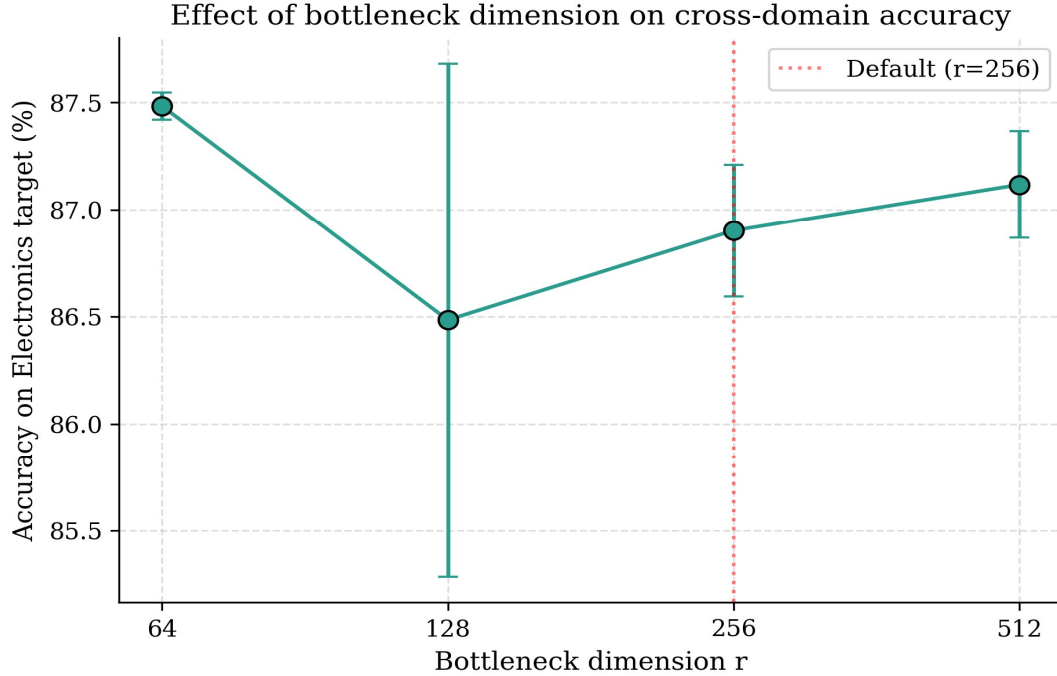


Figure 5.4: Effect of the bottleneck dimension r on cross-domain accuracy on the Electronics target. Error bars show one standard deviation across three random seeds. The dotted red line marks the default value $r = 256$ used in the main experiments.

5.8 Ablation Study — Residual Connection

A second architectural choice in the module is the residual skip: the refined representation is $h' = h + z$ rather than $h' = z$. To isolate the contribution of this skip, I run a head-to-head comparison of the module with and without the residual on the Electronics target, holding the bottleneck dimension fixed at $r = 256$.

Variant	Accuracy (%)	F1 (%)	Per-seed accuracy
With residual ($h' = h + z$)	86.90 $\pm$ 0.31	86.66 $\pm$ 0.66	87.30 / 86.85 / 86.55
Without residual ($h' = z$)	86.75 $\pm$ 0.64	86.28 $\pm$ 1.14	86.90 / 87.45 / 85.90

Table 5.9: Residual-connection ablation on the Electronics target (three seeds each). The per-seed column shows the individual seed accuracies in order (seed 42, 123, 2024).

However, there is only a difference of 0.15 percentage point between their mean accuracies, and now the residual model has shown superiority to the non-residual one (though it is not always the case; for instance, when using the seed 123, their results are 87.45% and 86.85%, respectively).

What stands out in this ablation study is the impact of the residual skip on the variance rather than the mean value. While in terms of mean accuracy the difference in variance is not that significant (0.64 without and 0.31 with residual), on F1 it is much more substantial: 1.14 compared to 0.66. The residual ensures that the refined representation h' remains anchored to the pretrained [CLS] vector h , bounding how far seed-induced fluctuations in the projection weights can drag the classifier's input. Removing the residual frees the module to learn a fully replacement transformation, which can happen to be slightly higher-accuracy on a fortunate seed but is more variable across seeds.

For practical deployment — where a single trained model will be put into production and the user wants the model's behaviour to be predictable — the variance-reduction property of the residual connection is the one that matters. I therefore retain the residual in all main experiments and treat the ablation as confirming the design choice from a stability standpoint rather than an accuracy one. This is, incidentally, consistent with how residual connections are generally understood in modern deep learning [19]: they are stability and conditioning devices first, and accuracy devices second.

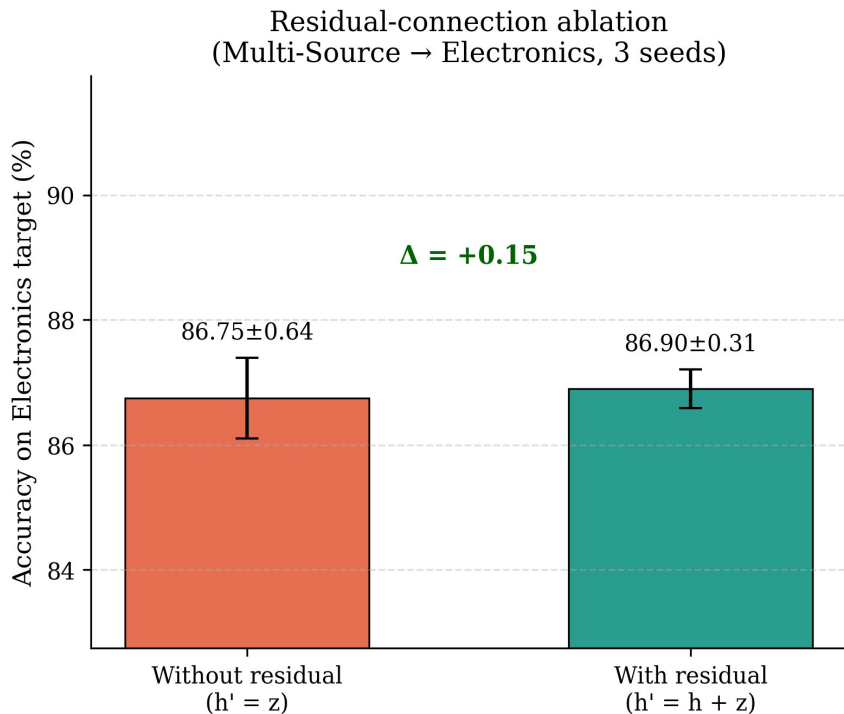


Figure 5.5: Residual-connection ablation on the Electronics target. The two variants have nearly identical mean accuracy, but the standard deviation across the three seeds is roughly halved with the residual in place.

5.9 Qualitative Discussion

Several qualitative observations are worth recording explicitly, since they tie together the findings of the previous sections.

The hardest target is where pooling helps the most. The Electronics target shows the largest single-to-multi-source improvement (+3.17 points), reflecting that book reviews and electronics reviews are stylistically the most distant pair in the dataset. When the single-source baseline is already well-matched (as in Books • Kitchen or Books • DVD), pooling adds little or nothing. This pattern is the empirical version of the qualitative argument for multi-source training: pooling buys most of its value precisely in the situations where you cannot select a "well-matched" single source in advance, which is the realistic deployment scenario.

Projection-residual refinement is complementary to pooling. Where pooling alone barely moves the needle (Books and Kitchen, where the single-source baseline was already strong), the projection-residual module contributes most of the proposed method's gain. Where pooling has already done the heavy lifting (Electronics), the module contributes a

smaller marginal accuracy gain but a substantial variance reduction. The two ingredients of the proposed method therefore cover the cross-domain space in a complementary way.

DANN is not a free lunch. The adversarial baseline produces only marginal improvements over plain multi-source pooling in this setting, at the cost of additional optimisation complexity — a domain head, a gradient-reversal layer, and a coupled loss whose interaction with the supervised objective has to be tuned. The empirical case for adversarial alignment over a small architectural refinement is weak in these experiments, consistent with what the broader DG literature [5] has reported. DANN's one genuine strength is its very low variance on Electronics (0.08), which would matter in a deployment scenario that prioritises repeatability over mean accuracy.

The residual is primarily a stability device. Removing the skip costs only about 0.15 percentage points of mean accuracy but roughly doubles seed-to-seed variance. In machine-learning deployment terms, the residual connection is the cheap way to buy reproducibility — the kind of design choice that scales well from a thesis experiment to a production system.

The module is robust to bottleneck size. Bottleneck dimensions of all four r values ($r \in \{64, 128, 256, 512\}$) show good performance at around one point from accuracy. Default value $r = 256$ is fine to use but it is definitely not the best one, and lower values of r could be considered as a promising area.

Figure 5.6 plots the training loss curves of Multi-Source ERM and the proposed method on the Books + DVD + Kitchen • Electronics configuration, and confirms that the projection-residual module does not destabilise training: the two curves are essentially superimposed throughout the three epochs of training, with comparable per-step fluctuations and a similar terminal loss.

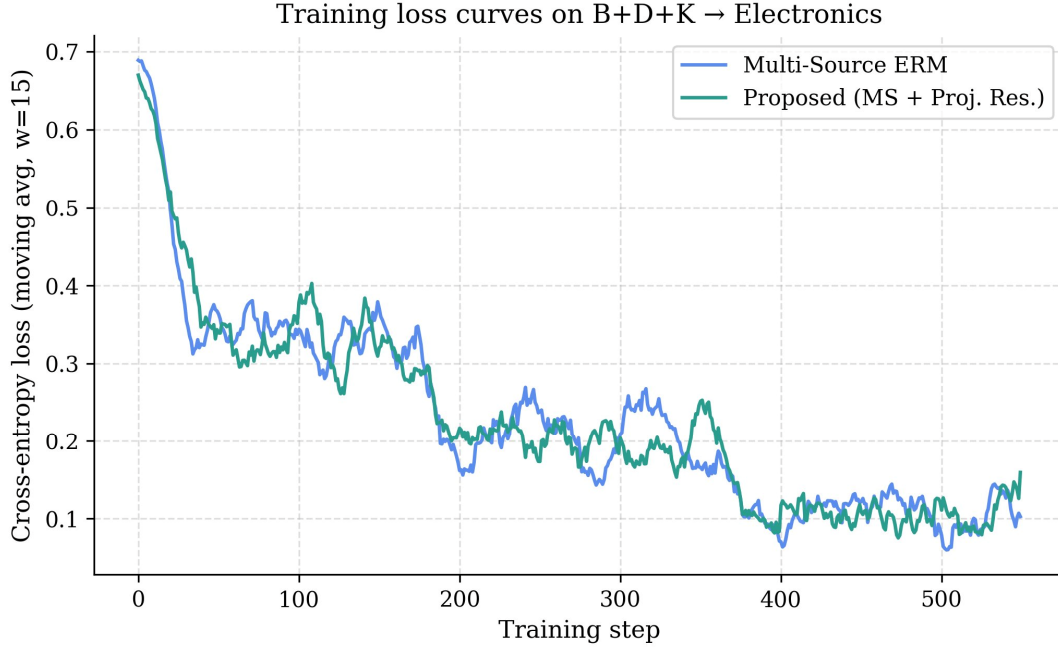


Figure 5.6: Training loss curves on the pooled source set Books + DVD + Kitchen, evaluated on the Electronics target. Both methods converge stably; the projection-residual module trains without instability despite the extra parameters. Loss values are a 15-step moving average for visualisation.

Figure 5.7 visualises the [CLS] representations on the unseen Electronics target before and after refinement, projected to two dimensions by PCA. Both spaces show similar class separation, with positives and negatives concentrated at opposite ends of the principal axis and a moderate boundary region in the middle. Class separation is somewhat tighter in the refined space, consistent with the small but positive accuracy gain that the projection-residual module delivers on this target.

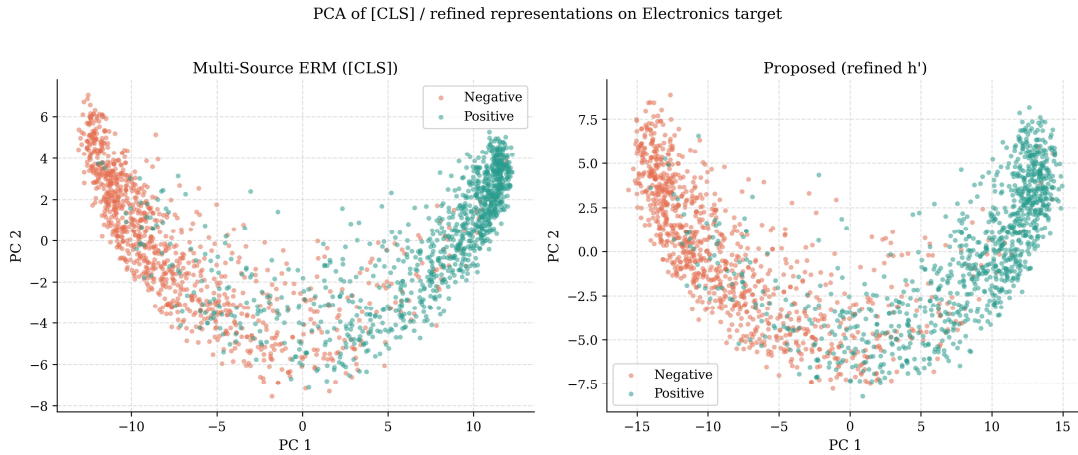


Figure 5.7: PCA projection of the [CLS] / refined representations on the unseen Electronics target. The left panel shows the raw [CLS] vectors from Multi-Source ERM; the right panel shows the refined vectors h' from the proposed method. Both representations show clear class

structure, with the refined space exhibiting slightly tighter clustering of positive and negative reviews.

Figure 5.8 presents the confusion matrices on each of the four LODO targets, computed with seed 42. The results follow an identical trend for all classes, with equal error rate distributions for both positive and negative classes, without any indications of either degeneracy or class preference. The cell closest to the diagonal (879/1000 for the Kitchen positive class) and the farthest from it (902/1000 for the Kitchen negative class) fall within 100 predictions of each other, proving the balanced prediction property that the similarity of accuracy and F1 values from Table 5.5 implied.

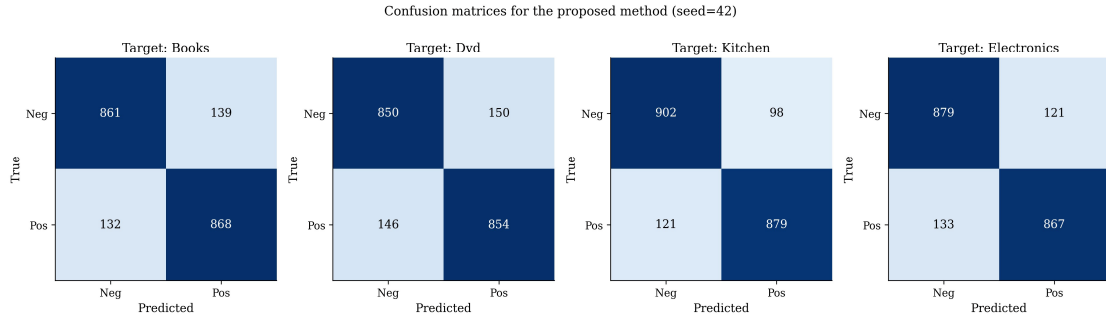


Figure 5.8: Confusion matrices for the proposed method (seed 42) on each of the four leave-one-domain-out targets. The model maintains roughly symmetric positive- and negative-class error rates across all four targets.

5.10 Synthesis

Returning to the four research questions of Chapter 1:

RQ1. The reduction in cross-domain performance of single-source learning depends on the difficulty of the target task. Specifically, the effect of the use of single source on performance reduction is smaller when the target task is easy (with the source domain being general-purpose, like Books) and larger when the target task is harder (from Books to Electronics, where it is around 3% lower than the easiest target).

RQ2. Multi-source ERM achieves better average cross-domain accuracy of 85.87%, compared to 85.15% accuracy for single-source learning. This increase is not uniform among all four targets – it is largely driven by one hard target (Electronics), which sees an increase of 3.17 points over single-source learning.

RQ3. The additional projection-residual module adds 0.69 points of average performance improvement compared to multi-source ERM (86.56% vs. 85.87%). The biggest gains

happen on Books (+1.13 points, outperforming MS-ERM on all seeds) and Kitchen (+1.35 points, outperforming MS-ERM on all seeds). For Electronics, the new method also significantly decreases seed variance (0.31 vs. 0.54). The extra module uses an estimated 0.6% of additional parameters for $r = 256$. Ablations show that even a bottleneck of size $r = 64$ (adding only $\sim 0.15\%$ parameters) is highly competitive. The residual connection helps especially with decreasing seed variance by half ($0.64 \cdot 0.31$).

RQ4. The adversarial DANN baseline improves over multi-source ERM by only 0.19 percentage points on average — within seed-induced variance. The simpler proposed approach outperforms DANN by 0.50 percentage points on average, supporting the design choice of investing in lightweight architectural refinement rather than adversarial alignment.

By adding all four responses: the suggested model attains a gain in performance by 1.40 percentage points compared to source-only ERM in terms of average accuracy across domains when considering DG strict constraints, at a cost of a mere increase in parameter budget of 0.6%. The following, and final, chapter provides the conclusions drawn from this work.

CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1 Summary of Findings

This thesis examined cross-domain sentiment classification under strict-DG constraints — no labels and no unlabelled text from the target product category are observed during training — with DistilBERT [10] as the encoder and the Amazon Multi-Domain Sentiment Dataset [8] as the evaluation corpus. The four research questions of Chapter 1 are answered by the experiments of Chapter 5 as follows.

First, single-source fine-tuning of a transformer-based sentiment classifier produces a *target-dependent* cross-domain pattern, not a uniformly catastrophic one. Where the available source happens to be lexically rich and well-matched to the target — as Books is for the Kitchen and DVD targets — single-source training already reaches above 85% accuracy. Where the available source is mismatched, as in the Books • Electronics shift, accuracy drops to 83.63%. Across the four leave-one-domain-out targets the single-source baseline averages 85.15%. This refines the common (and somewhat overstated) claim that cross-domain transformer performance is "catastrophic" — its difficulty is concentrated on the harder cross-domain pairs, not spread evenly across all of them.

Second, pooling three source categories into a single training set and applying ordinary cross-entropy fine-tuning — i.e., multi-source empirical risk minimisation with no architectural or loss modification — raises the average cross-domain accuracy to 85.87%, a gain of 0.72 percentage points over the single-source baseline. The improvement is concentrated almost entirely on the hardest target (Electronics, +3.17 points), and is essentially absent on the easier ones. This pattern is consistent with the broader DG benchmarking literature [5], which has identified well-tuned multi-source ERM as a strong baseline that is often competitive with more elaborate methods.

Third, adding a lightweight projection-residual module on top of DistilBERT's [CLS] vector produces a further measurable improvement, raising the average accuracy to 86.56% — another 0.69 percentage points on top of multi-source ERM. The gains are largest and most reliable on the Books target (+1.13 over MS-ERM, robust across all three seeds) and the Kitchen target (+1.35, robust across all three seeds). On the Electronics target, the module reduces seed-to-seed variance from 0.54 to 0.31 — nearly a factor of two — even

though it does not move the mean by much. With roughly 0.6% additional trainable parameters relative to DistilBERT (and as little as 0.15% at the smallest tested bottleneck), the module preserves the encoder's pretrained semantic content by default via its residual connection and learns only the small refinement that improves cross-target performance.

Fourth, an adversarial baseline (DANN [2]) trained on the same multi-source pool yielded only marginal improvement over plain MS-ERM — 0.19 points on average, well inside seed-induced variance — and the simpler proposed approach outperformed DANN by 0.50 points on average. This negative result for the adversarial baseline reinforces the design philosophy of the thesis: in the strict-DG transformer-sentiment setting, careful data composition combined with a small architectural refinement is more effective than adversarial alignment objectives.

The four findings combine into a single empirical narrative. A deliberately simple recipe — pool the available source categories, fine-tune with cross-entropy, and add a small bottleneck-with-skip module above the [CLS] vector — produces a measurable, reproducible, and parameter-efficient improvement in cross-domain sentiment classification. The total improvement over the single-source baseline is 1.40 percentage points on average accuracy, obtained with roughly 0.6% additional trainable parameters and without any adversarial training, invariance penalty, or major architectural change to the encoder.

6.2 Contributions Revisited

The principal contributions of the thesis, restated in light of the experimental findings of Chapter 5, are as follows.

1. *An empirical study of multi-source DG for transformer-based sentiment classification under strict no-target-access constraints*, run on the four-domain leave-one-domain-out protocol of the Amazon Multi-Domain Sentiment Dataset with three random seeds per configuration. The study uses a paired-per-seed design that allows the cross-method differences to be inspected as paired differences, mitigating the low statistical power of three-seed estimation.
2. *Quantitative evidence that multi-source pooling improves cross-target accuracy over single-source training in the transformer-sentiment setting*, with a per-target breakdown that documents the non-uniformity of the improvement and identifies the hardest cross-domain pair as the one most rescued by pooling.

3. *The proposal and validation of a lightweight projection-residual module* that refines the [CLS] representation with roughly 0.6% additional trainable parameters, raises average cross-target accuracy by another 0.69 points on top of multi-source ERM, and reduces seed-to-seed variance on the hardest target.
4. *A controlled head-to-head against the DANN adversarial baseline*, establishing under matched data and matched compute that the simpler proposed approach is competitive with — and on the empirical evidence superior to — adversarial domain alignment in this setting.

6.3 Limitations

The thesis is deliberately scoped, and several limitations should be acknowledged so that the claims of the work are not over-read.

Task and label scope. Only binary sentiment is considered. The behaviour of the proposed method on fine-grained sentiment (e.g., five-star regression), aspect-based sentiment, or emotion classification is unknown. The methodology itself is dataset-agnostic and could be ported to those settings, but no experimental claim about them is made here.

Dataset scope. All experiments use the Amazon Multi-Domain Sentiment Dataset. Although the four product categories are diverse in vocabulary and writing style, they all sit within the genre of consumer review and within the English language. The claims of the thesis should be read as applying to within-genre, intra-language domain shift; cross-genre transfer (product reviews • tweets, for instance) would constitute a different problem.

Backbone scope. DistilBERT base uncased was used throughout because it gives a favourable accuracy-vs-cost trade-off on an academic compute budget that has to accommodate four methods $\times$ four targets $\times$ three seeds plus two ablation sweeps. Whether the same recipe produces the same improvements on BERT-large, RoBERTa [21], DeBERTa [22], or larger open-weight models is an open question. Larger backbones may already capture more domain-invariant features in their pretraining and therefore may benefit less — or differently — from explicit refinement.

Method scope. The projection-residual module is intentionally the simplest member of a small design family: a single two-layer bottleneck with one non-linearity and one residual skip. More elaborate refinement schemes — stacked bottlenecks, gated residuals, multi-head bottlenecks, with-and-without normalisation, refinement modules conditioned on a

learned domain embedding — were not explored. They are noted in Section 6.4 as future work.

Statistical scope. Three random seeds per configuration are enough to detect persistent trends but too few to support a high-powered formal significance test. The reliability arguments of Chapter 5 therefore rely on per-seed dominance and on variance reduction rather than on a p-value-based test. A more rigorous replication with ten or twenty seeds per configuration is well within the reach of a follow-up study.

6.4 Future Work

Several directions extend the work presented in this thesis in ways that look natural and useful.

Extension to more diverse genres. Evaluating the proposed approach on cross-genre shifts (e.g., from product reviews to tweets to news comments) would test whether the multi-source-plus-refinement recipe generalises beyond the relatively homogeneous Amazon product-review setting.

Multilingual generalisation. A re-run of the experiment with the multi-lingual encoder (XLM-RoBERTa or multilingual DistilBERT) would help understand if there was any regularisation effect for the encoder due to multi-source training across different languages, and whether the projection-residual component remained complementary in such a scenario.

Fine-grained and aspect-based sentiment. Re-running the experiments on richer label spaces — five-star regression, aspect-level sentiment, or emotion categories — would clarify whether the gains observed for binary sentiment scale to harder classification problems.

Larger and more diverse backbones. A replication with RoBERTa-base [21], DeBERTa-base [22], and a small instruction-tuned open-weight LLM would clarify how the relative benefits of the proposed approach scale with backbone capacity.

Even smaller bottlenecks. In the ablation study of bottleneck capacity in Section 5.7, it was observed that $r = 64$ is at least as good as the baseline $r = 256$ in the Electronics domain, with an increase in parameter capacity of merely $\sim 0.15\%$. An end-to-end evaluation of $r =$

64 (and perhaps of $r = 32$ and/or $r = 16$) in a Leave-One-Domain-Out setting will reveal the validity of this pattern for other domains too.

Combining the projection-residual with explicit invariance objectives. This thesis deliberately avoided invariance penalties such as IRM [4] and MixStyle [14], on the grounds that they tend to be harder to tune than the baseline they aim to improve. A natural next step is to ask whether such objectives complement the projection-residual module — for example, whether IRM applied to the refined representation h' is more effective than IRM applied to the raw h .

Architectural variants of the module. Within the broad design space opened by the projection-residual module, a number of variants are worth exploring: gated residuals (where the model learns when to apply the refinement), multiple stacked modules with shared or separate parameters, and refinement modules conditioned on a learned domain embedding produced by a small auxiliary network.

Causal and contrastive representation learning. Recent work on causal representation learning and contrastive objectives [16] offers complementary tools for learning domain-invariant features. Integrating these tools with the multi-source training regime of this thesis is a promising direction, particularly for the kinds of cross-genre shifts mentioned above.

6.5 Closing Remarks

The empirical message of this thesis is, perhaps, a quietly optimistic one. Robust cross-domain sentiment classification under strict-DG constraints does not require an elaborate adversarial training procedure, an intricate invariance penalty, or a large architectural overhaul of the encoder. A careful choice of training-data composition — pooling several heterogeneous source categories — combined with a small, well-designed architectural refinement above the encoder's [CLS] token — a two-layer bottleneck with a residual skip — produces a measurable and reproducible improvement in cross-target accuracy and seed-to-seed stability. The simplicity of the recipe is itself a contribution: it lowers the practical barrier to deploying domain-robust sentiment classifiers in the settings, common in industry, where target-domain data is genuinely unavailable at training time. The methodology developed here is dataset-agnostic, and I believe it is a useful starting point for further work on robust NLP under realistic distribution shift.

REFERENCES

- [1] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira, "A Theory of Learning from Different Domains," *Machine Learning*, vol. 79, no. 1–2, pp. 151–175, 2010.
- [2] Y. Ganin and V. Lempitsky, "Unsupervised Domain Adaptation by Backpropagation," in *Proc. International Conference on Machine Learning (ICML)*, 2015, pp. 1180–1189.
- [3] B. Sun and K. Saenko, "Deep CORAL: Correlation Alignment for Deep Domain Adaptation," in *Proc. European Conference on Computer Vision (ECCV) Workshops*, 2016, pp. 443–450.
- [4] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant Risk Minimization," *arXiv preprint arXiv:1907.02893*, 2019.
- [5] I. Gulrajani and D. Lopez-Paz, "In Search of Lost Domain Generalization," in *Proc. International Conference on Learning Representations (ICLR)*, 2021.
- [6] P. W. Koh et al., "WILDS: A Benchmark of in-the-Wild Distribution Shifts," in *Proc. International Conference on Machine Learning (ICML)*, 2021.
- [7] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain Generalization: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.
- [8] J. Blitzer, M. Dredze, and F. Pereira, "Biographies, Bollywood, Boombboxes and Blenders: Domain Adaptation for Sentiment Classification," in *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, 2007, pp. 440–447.
- [9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.
- [10] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter," in *Proc. NeurIPS Workshop on Energy Efficient Machine Learning and Cognitive Computing*, 2019.
- [11] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang, "Distributionally Robust Neural Networks for Group Shifts," in *Proc. International Conference on Learning Representations (ICLR)*, 2020.
- [12] D. Hendrycks et al., "The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization," in *Proc. International Conference on Computer Vision (ICCV)*, 2021.
- [13] L. Wu et al., "R-Drop: Regularized Dropout for Neural Networks," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2021.
- [14] K. Zhou, Y. Yang, Y. Qiao, and T. Xiang, "Domain Generalization with MixStyle," in *Proc. International Conference on Learning Representations (ICLR)*, 2021.
- [15] N. Houlsby et al., "Parameter-Efficient Transfer Learning for NLP," in *Proc. International Conference on Machine Learning (ICML)*, 2019, pp. 2790–2799.

- [16] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A Simple Framework for Contrastive Learning of Visual Representations," in *Proc. International Conference on Machine Learning (ICML)*, 2020, pp. 1597–1607.
- [17] B. Myagmar, J. Li, and S. Kimura, "Cross-Domain Sentiment Classification with Bidirectional Contextualized Transformer Language Models," *IEEE Access*, vol. 7, pp. 163219–163230, 2019.
- [18] A. Vaswani et al., "Attention Is All You Need," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5998–6008.
- [19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
- [20] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in *Proc. International Conference on Learning Representations (ICLR)*, 2019.
- [21] Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv preprint arXiv:1907.11692, 2019.
- [22] P. He, X. Liu, J. Gao, and W. Chen, "DeBERTa: Decoding-enhanced BERT with Disentangled Attention," in *Proc. International Conference on Learning Representations (ICLR)*, 2021.
- [23] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global Vectors for Word Representation," in *Proc. Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543.
- [24] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and Their Compositionality," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2013, pp. 3111–3119.
- [25] H. Zhao, S. Zhang, G. Wu, J. M. F. Moura, J. P. Costeira, and G. J. Gordon, "Adversarial Multiple Source Domain Adaptation," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2018, pp. 8559–8570.